METRICS OF GENETIC RELATEDNESS IN APPLICATIONS OF HUMAN GENOMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Haley Selma Hunter-Zinck

August 2014

METRICS OF GENETIC RELATEDNESS IN APPLICATIONS OF HUMAN GENOMICS

Haley Selma Hunter-Zinck, Ph.D.

Cornell University 2014

Measuring genetic relatedness is fundamental to many applications of human genomics. Genetic relatedness can be defined in several different ways ranging from global, genome-wide estimations to confined, locus-specific measurements. Local relatedness is often represented as identity-by-descent (IBD), but IBD is an approximation of time to most recent common ancestor (TMRCA), and using TMRCA directly has the potential to be a more informative metric. Applications using metrics of genetic relatedness in human genomics include diverse topics as inbreeding and relatedness measurements, population structure, demographic history, effective population size estimates, haplotype phasing, genome-wide association studies, natural selection inference, and many others. In this dissertation, I will describe three projects using metrics of human relatedness in three different applications. I will first give a general overview of the definition and use of metrics of genetics relatedness and set the context in applications in the field of human genomics. I will then describe a project using IBD to look at population substructure among a sample of Qatari individuals. In the subsequent two chapters I will move on to using TMRCA to develop more general methods of natural selection inference and association mapping. The last two chapters provide evidence that TMRCA provides a more informative and unifying metric than IBD for two different applications in human genomics. IBD was a metric of choice because of necessity. However, with the production

of high coverage whole-genome sequences and advancement of computational methodology, using TMRCA as a metric of genetic relatedness is now feasible, providing an avenue to further biological insights via this more informative metric.

BIOGRAPHICAL SKETCH

Haley Hunter-Zinck was born in Los Angeles, California and studied bioengineering: bioinformatics at the University of California, San Diego (UCSD) from 2005-2009, graduating with a bachelor of science. While at UCSD, she participated in research in two laboratories, the James Nieh laboratory studying the ecology and communication of bees and the Trey Ideker laboratory studying correlations between chromatin marks and gene expression. She also participated in research internships abroad for two summers, the first in Brazil in 2007 and the second in Australia in 2008. In Brazil, she worked as a field assistant to a Nieh lab Ph.D. student, Elinor Lichtenberg, studying eavesdropping behaviors of stingless bees in University of São Paulo in Ribeirão Preto. The next summer, in Melbourne, Australia, she participated in a computational biology internship at Monash University working jointly for Dr. Philip E. Bourne at UCSD and Dr. David Abramson at Monash.

In 2009 Haley was accepted into the Tri-Institutional Program in Computational Biology and Medicine at Cornell. From 2009-2010 she rotated with Dr. Andrew Clark at Cornell University in Ithaca, Dr. Jason Mezey at both Cornell University and the Weill Cornell Medical School, and Dr. Robert Klein at the Memorial Sloan Kettering Cancer Center. During her rotations, she worked on a variety of projects including Qatari genetic population structure, penalized regression methods for association studies, and natural selection on genes involved in cancer. She joined the Clark lab in 2010 and began working on identity-by-descent but gradually transitioned to method development projects involving coalescent statistics such as time to most recent common ancestor. She is now transitioning to the field of medical informatics using the computational skills she acquired at Cornell to solve applied problems in medicine.

To my family

and in memory of my mother,

Dr. Jennifer Joanne Zinck

(1955-2011)

ACKNOWLEDGEMENTS

The completion of this work would not have been possible without the help and support of many generous people. I would like to thank my adviser, Dr. Andrew G. Clark, and my committee members Dr. Alon Keinan, Dr. Robert Klein, Dr. Jason Mezey, and Dr. Adam Siepel. I would further like to acknowledge members of the Clark lab including Roman Arguello, Nancy Chen, Tim Connallon, Angela Early, Jen Grenier, Yu (Amanda) Guo, Keegan Kelsey, Amanda Manfredo, Margarida Cardoso Moreira, Srilakshmi Raj and Cris Van Hout; members of the Keinan lab including Leonardo Arbiza, Paul Billing-Ross, Feng Gao, Elodie Gazave, Aviv Madar, Aaron Sams, and Andrea Slavney; and members of the Mezey lab including Francisco Agosto Pérez, Gabriel Hoffman, Pavel Korniliev, Ben Logsdon, Larsson Omberg and Monica Ramstetter for help and discussion throughout my Ph.D. It has been a pleasure working with such motivated, intelligent, and creative people.

I would also like to acknowledge that this work was supported by the NIH Training Grant 1T32GM083937, the Tri-Institutional Training Program in Computational Biology and Medicine, and the NIH R01 HG003229 Grant.

A big thanks to my running partners, Roman Arguello, Nancy Chen, Keegan Kelsey, Ben Logsdan, Pavel Korniliev, Amanda Manfredo, Larsson Omberg, and Nadia Singh, who have helped keep me sane throughout the last five years. Further, I would like to acknowledge the continuous help, support, and encouragement from my close friends and colleagues Diana Chang and Jaaved Mohammed.

I would also like to thanks my parents for moral and financial support throughout my undergraduate career, which brought me to Cornell. And finally to my husband, Licurgo Benemann de Almeida, for his unfaltering love

and support.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The concept of genetic relatedness is fundamental to many applications of human genomics. In particular, genetic relatedness is crucial to applications of population genetics and the association of genetic mutations with simple and complex phenotypes. Genetic relatedness can define relationships between individuals on both a genome-wide and a local scale, and the concept is useful in many applications on both levels. Although many metrics of genetic relatedness exist, identity-by-descent (IBD) and time to most recent common ancestor (TMRCA) are two metrics used in a diverse set of applications. Genetic sequences are considered identical-by-descent if they originate from a common ancestor at a time in the recent past. In contrast, TMRCA defines the explicit time at which the common ancestor of the sequences arose, rendering IBD a binary approximation of TMRCA. As sequence data becomes more prevalent, previously underutilized metrics of genetic relatedness, such as TMRCA, become possible to calculate with greater precision, and the continuing use of approximations of relatedness, such as IBD, in samples of putatively unrelated individuals is called into question. Here, I will first present examples of past and developing research showing the generality of relatedness metrics across different applications in human genomics and then argue that TMRCA provides a more informative metric than IBD for many of these applications.

In population genetics, genetic relatedness establishes the groundwork for understanding populations and their relationships. Wright first defined effective population size, the inbreeding effective size, in terms of relatedness as the change in the probability of identity by descent through successive generations

[1]. Analyzing population genetic structure is another application that relies on genetic relatedness because the analysis, in essence, determines the pattern of relatedness between individuals at a genome-wide scale [2, 3, 4]. Individuals within reproductively isolated populations are more related to each other, on average, than to individuals from different populations, and in this way, populations form genetic clusters consisting of related groups that can be visually represented by methods such as principal components analysis. Furthermore, demographic inference uses relatedness, either explicitly or implicitly, to reconstruct population history as the distribution of relatedness between individuals in a population sample is reflective of their past demographic history [5, 6, 7, 8, 9]. For example, a severe bottleneck reduces the effective population size of a population, making individuals of subsequent generations more related to each other than individuals from the same population before the bottleneck. Finally, natural selection methods also utilize relatedness to assess the presence of selection [10, 11, 12]. Excess relatedness in a region compared to the null indicates that forces other than demography are acting on a locus, as when, for example, regions under the influence of positive selection show an increase in local relatedness between individuals after a beneficial haplotype swept through the population. The examples above are just a sample of the applications demonstrating that the field of population genetics relies heavily on the concept of genetic relatedness.

Understanding the population genetics concepts described above as products of genetic relatedness is fundamental to the association of mutations with disease or other phenotypes. Genome-wide association studies (GWAS) require knowledge of population structure and demography in order to properly construct association models and avoid spurious associations [13]. Additionally,

inference of natural selection can help to interpret and support statistical associations of mutations with Mendelian and complex disease such as, for example, in the NOD2 locus for Crohn's disease [14]. But most directly, GWAS methods themselves use statistics derived from sharing local ancestry directly to associate variants with phenotypes as in an admixture mapping study by Pasaniuc, *et al.* that investigated the genetic factors underlying coronary heart disease and type 2 diabetes [15]. Genetic relatedness is an essential concept to many applications of human genomics.

In many of the previously mentioned applications, genetic relatedness is expressed in terms of identity by descent (IBD). The definition of IBD varies depending on the context of the sample in question [16]. In pedigrees, IBD is defined with respect to founders who are either unrelated or related with known relatedness coefficients. In populations, IBD is defined with respect to a chosen threshold in time for which individuals are IBD at a locus if they share a common ancestor more recently in the past than the given threshold. In this way, methods identify individuals who share a common ancestor in the recent past for a particular region of their genome. However, in practice, IBD inference methods have more power to detect long IBD segments and miss short IBD segments no matter how recently in the past the common ancestor occurred. But even if the metric does not fully capture the relationship between two individuals, IBD does provide a valuable approximation of the relatedness between two samples at a locus.

Previous studies have used IBD in applications of population genetics and beyond. Recently developed methods have assessed population genetic structure using haplotype similarity, a proxy for IBD, and applied these methods

successfully to large and diverse samples such as the Human Genome Diversity Panel [4]. Several natural selection inference methods also used IBD to investigate the effects of selection in the HapMap 3 data set genotyped samples [10, 17]. IBD and population level ancestry sharing have also been applied to associate mutations with complex diseases such as benign recurrent intrahepatic cholestasis, coronary heart disease, type 1 diabetes, type 2 diabetes and breast cancer [18, 15, 19]. IBD has been used as a metric of genetic relatedness in many applications of human genomics.

However, there are several drawbacks to using IBD as a metric of genetic relatedness when analyzing population samples of distantly related individuals. The definition of IBD in population samples requires the selection of an arbitrary threshold of time in the past, which can be explicitly defined but is often implicitly stipulated by the IBD inference method's power. Methods cannot identify shared regions smaller than roughly 1 cM because of the confounding effect of linkage disequilibrium, and IBD regions are expected to be smaller than this threshold after the common ancestor originates approximately 100 generations in the past. Furthermore, the binary nature of IBD in most inference methods, although multiple IBD state models are possible, is not as informative as a continuous metric. All these factors limit the utility of IBD.

Because of the difficulties mentioned above, the application of IBD in human genomics has made notable but narrow progress. Natural selection inference using IBD did not initially reveal any new loci, although more recent methods have had more success [10, 17]. But despite recent advancement, IBD is only applicable to recent positive selection, while selection in other time periods or other modes of selection would still be difficult to detect with IBD. Moreover,

IBD-mapping has only proved applicable for genetic architectures of disease that depend on an extreme level of allelic heterogeneity at a single locus [19]. Although models of allelic heterogeneity might be more common with recent super-exponential growth and the excess of rare variation, the scope of IBD-mapping is still confined to specific genetic architectures of disease.

As human genetics enters the age of sequence data, methods must take full advantage of the additional information data provide. Most IBD inference methods were designed in the genotyping era, although more recently benchmarking studies and adaptations of these methods for sequencing data have been conducted [20, 21, 22, 17, 23, 24]. Despite these adaptations, IBD inference methods still do not take full advantage of sequencing data. Even though more accurate inference of IBD blocks is helpful, the biological variance in IBD block length due to meiotic recombination, limits the power of inference of timing of most recent common ancestors [16]. In other words, even with perfect inference of IBD blocks between two individuals, the exact relationship between the two individuals is still uncertain. Although much has been accomplished using IBD as a metric of genetic relatedness, potentially more informative metrics of genetic relatedness can be derived from whole-genome sequences.

Studies have attempted to surpass the approximations of IBD with construction of the full ancestral recombination graph, which is more informative for addressing applications in human genomics and many other species [16]. However, these methods, although improving, are still too computationally intensive for large samples [25]. Fortunately, although not ideal, recent methods infer pairwise time to most recent common ancestor (TMRCA) in a scalable fashion, which advances one step beyond IBD [5]. Although not providing the full infor-

mation of the ancestral recombination graph, pairwise TMRCA values provide a continuous metric, instead of a binary metric based on an arbitrary threshold, to measure local relatedness between individuals.

In the following dissertation, I will describe three projects using metrics of genetic relatedness to tackle different problems in population genetics and association mapping. While I will discuss some results using IBD, I will also describe two projects using TMRCA, a potentially more informative metric.

In the second chapter, I will describe a project investigating population genetic structure for a sample of individuals from the country of Qatar in the Middle East [26]. Principal components analysis of the sample reveals three subpopulations corresponding to ancestral populations from Africa, the Middle East, and Asia. Furthermore, each sub-population shows a different degree of consanguinity, within individual IBD, and other population genetic metrics, reflecting the different population history and demographics of each subgroup. Genetic relatedness serves as a valuable concept to deconstruct the population structure and population history of this sample.

In the next chapter, I will describe a method for natural selection inference using TMRCA in a diverse sample. While there already exist a wide variety of genome-wide inference methods for natural selection, all of these methods target a specific mode of selection and scale of time [27, 14, 28]. Natural selection is known to distort local genealogies among a sample of individuals, so I will utilize distributions of pairwise TMRCA values as an approximation of local genealogies to see which loci are distorted compared to neutral genealogies. I will then show that this method tackles a larger scope of natural selection modes in a larger range of ages and strengths than previous methods, including those

utilizing IBD.

In the final chapter, I will investigate the potential of using TMRCA in GWAS. The missing heritability problem is still rampant, and one factor hypothesized to contribute to this problem is the effect of rare variants [29, 30]. Many of the new association methods address rare variant disease models but only work well on the genetic architectures for which they were designed [31]. Just as distortions in local genealogies, as assessed by TMRCA, create a general model for natural selection inference, I will show how using pairwise TMRCA values are a more informative metric for GWAS as well, but especially in the area of rare variant associations. Intuitively, in loci underlying the expression of a disease or phenotype, individuals with similar phenotypic values should be more related to one another than individuals with highly disparate phenotypes. Even if multiple rare variants affect the phenotype, causal loci should exhibit multiple clusters of related individuals that non-causal loci do not. TMRCA provides an explicit and precise metric for measuring this genetic similarity between individuals in this framework.

The concept of genetic relatedness is fundamental to our understanding of human genomics, and IBD has been an important metric of genetic relatedness. I will show how using TMRCA greatly increases and expands our understanding of the changes in genetic structure. In this way, I hope to promote the use of TMRCA as a metric of human relatedness in future applications of human genomics.

CHAPTER 2

POPULATION GENETIC STRUCTURE OF THE PEOPLE OF QATAR

Haley Hunter-Zinck,[1] Shaila Musharoff,[1] Jacqueline Salit,[2] Khalid A. Al-Ali,[3] Lotfi Chouchane,[4] Abeer Gohar,[4] Rebecca Matthews,[4] Marcus W. Butler,[2] Jennifer Fuller,[2] Neil R. Hackett,[2] Ronald G. Crystal,[2] and Andrew G. Clark[5,6]

2.1 Abstract

People of the Qatari peninsula represent a relatively recent founding by a small number of families from three tribes of the Saudi peninsula, Persia, and Oman, with indications of African admixture. To assess the combination of this founding effect and the customary first-cousin marriages among the ancestral Islamic populations on Qatars population genetic structure, we obtained DNA samples from 168 self-reported Qatari nationals sampled from Doha, Qatar and performed hybridizations to Affymetrix Genome-Wide Human SNP Array 5.0 to obtain genotype calls of nearly 500,000 single nucleotide polymorphisms (SNPs) in each individual. Principal components analysis was performed along with

[1]Program in Computational Biology and Medicine, Cornell University, Ithaca, NY 14850, USA
[2]Department of Genetic Medicine, Weill Cornell Medical College, New York, NY 10021,USA
[3]Department of Health Sciences, College of Arts and Sciences, Qatar University, Doha, Qatar
[4]Department of Genetic Medicine, Weill Cornell Medical College Qatar, Doha, Qatar
[5]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14850, USA
[6]Correspondence: Andrew G. Clark, ac347@cornell.edu, 607-255-0527

Reprinted from the American Journal of Human Genetics, Vol 87, Haley Hunter-Zinck, Shaila Musharoff, Jacqueline Salit, Khalid A. Al-Ali, Lotfi Chouchane, Abeer Gohar, Rebecca Matthews, Marcus W. Butler, Jennifer Fuller, Neil R. Hackett, Ronald G. Crystal, and Andrew G. Clark, Population Genetic Structure of the People of Qatar, Pages No. 17-25, Copyright (2010), with permission from Elsevier

samples from the Human Genetic Diversity Project (HGDP) dataset, revealing three clear clusters of genotypes whose proximity to other human population samples is consistent with Arabian origin, a more eastern or Persian origin, and individuals with African admixture. The extent of linkage disequilibrium (LD) is greater than that of African populations, and runs of homozygosity in some individuals reflect substantial consanguinity. However, the variance in runs of homozygosity is exceptional, and a population in which nearly half the population marries first cousins generally reflects greater identity-by-descent sharing than was observed. Despite the fact that the SNPs of the Affymetrix 500k chip were ascertained with a bias toward SNPs common in Europeans, the data strongly support the notion that the Qatari population could provide a valuable resource for the mapping of genes associated with complex disorders and that tests of pairwise interactions are particularly empowered by populations with elevated LD like the Qatari.

2.2 Introduction

The population of the State of Qatar is, like many modern societies, facing a growing threat from diabetes, obesity, and cardiovascular disease. Recent progress using genome-wide association studies (GWAS) has identified many additional genetic factors that appear to inflate the risk of disorders in some individuals [32, 33, 34, 35]. A drawback of the GWAS approach has been the limitation to largely people of European ancestry. Validation of risk factors identified in European GWAS can be conducted in different population samples and may often produce negative results. For example, while PPAR-gamma is associated with diabetes in some individuals of European descent, the gene is not a

risk factor in a Qatari population sample [36]. These results only further support the need to uncover non-European risk factors. A study of the population structure of the people of Qatar, as inferred by genetic testing, is necessary in order to determine how best to perform GWAS and other genetically-assisted analyses of risk in the Qatari population.

Based on surnames and oral history, it is thought that the bulk of the Qatari population originates from the Saudi peninsula, Persia, and Oman, with a minority descending from individuals of Africa and Southeast Asia. The people described as Arab are descendants of tribes from the Arabian Peninsula, including coastal tribes of pearl divers and the Hadar as well as the Bedouin nomads. The Ajam, or Iranian-Qatari, are descendants of merchants and craftsmen who migrated from Persia, and the majority of the Ajam speaks Farsi. Another group, the Abd, is descended from African slaves brought from Zanzibar via Oman to Qatar [37]. Qatars complex history makes the region especially interesting in determining whether population genetic methods of analysis reveal patterns of genetic polymorphism that are consistent with the country's history.

In keeping with the customs of Islam, first-cousin marriages have been widely accepted in Qatar and may have represented about half of all marriages in the region. More recent studies indicate that the rate of first-cousin marriages has fallen to about 22% but that attitudes toward consanguinity have remained accepting [38]. A high level of recurrent consanguinity would have a profound impact on the genetic structure of a population, as well as a distinct influence on the measures of population substructure. Here, we perform an analysis of high-density SNP genotyping chips on a sample of 168 individuals from the Qatari peninsula, and we attempt to reconcile the genetic information with the

historical understanding of this region.

2.3 Subjects and methods

2.3.1 Sample collection and SNP data collection

Human subjects were recruited under ongoing protocols approved by the Institutional Review Boards of Hamad Medical Corporation (#392/2006, #9093/09, #373/2006) and Weill Cornell-New York (#0605008516, #0904010340, #0604008489, #0806009874). All subjects had a general medical exam, basic demographic information and blood collected. DNA was extracted from blood using the Qiagen Blood kit and the DNA was quality controlled, requiring A260/A280 ratio of 1.8 - 2.1, quantified with a NanoDrop spectrophotometer. Frozen DNA, diluted to 50 ng/ml, was processed as recommended by Affymetrix Genome-Wide Human SNP Array 5.0. Processing involved restriction digestion, PCR amplification, purification and labeling. Aliquots were removed during processing to ensure the size profile and yield were within acceptable limits. After hybridization and washing the chips were scanned and quality control was performed with select heterozygous control SNPs. The Bayesian Robust Linear Model with Mahalanobis (BRLMM-P) algorithm was used to generate SNP calls and additional quality control was performed for call rates, consistency with self-reported ethnicity, relatedness to other samples and gender. Of the 168 initial subjects, we identified 156 unrelated subjects, defined as sharing less that 20% of their genome identical by descent (IBD) with any other individual as calculated via PLINK v1.06.8. Only the 156 unrelated

individuals were used in the following analysis.

2.3.2 SNP trimming for population structure inference

PLINK was used to prune the 440,794 SNPs down to 67,735 SNPs with a minor allele frequency greater than 5%, a missingness rate less than 1%, and a Hardy-Weinberg equilibrium deviation P-value of no less than 0.001 [39]. SNPs were pruned for pairwise linkage disequilibrium (r^2) maximum threshold of 0.5 using PLINKs –indep-pairwise command. We used the resulting subset of SNPs for the STRUCTURE analysis.

2.3.3 Inference of population clustering by STRUCTURE

The 67,735 SNPs were used in the program STRUCTURE to infer the clustering of individuals into affinity groups that behave like panmictic populations [40]. We applied this program assuming two, three, four, and five subpopulations on separate runs with 10,000 burn-in iterations and 10,000 iterations after burn-in. To determine the most likely number of subpopulations, we used the likelihood score calculated within the STRUCTURE program and the recommendations listed in the software documentation.

2.3.4 Selection of a reference sample and SNP filtering for remaining analysis

In order to assess existing population structure and standard population genetic parameters, we performed several forms of analysis on the Qatari sample with the Human Genome Diversity Project (HGDP) sample data as a reference [41]. We chose the HGDP sample because the data provides genotypes from populations around the globe, allowing us to construct an informed picture of how the Qatar population sample relates to other human population samples from several geographic regions. Each dataset was filtered to remove SNPs with minor allele frequency (MAF) less than 5% and an overall missingness greater than 1% as these results often indicate genotyping errors. We then filtered each Qatar group and HGDP population sample separately for Hardy-Weinberg equilibrium (HWE) deviations with P-value less than 0.001 in order to remove additional genotyping errors. By filtering each sample separately, we avoided eliminating SNPs deviating from HWE due to the Walund Effect, therefore retaining SNPs that deviate from HWE because of existing population structure and not simply genotyping errors [42]. Non-biallelic SNPs and unmapped SNPs were also removed. Because the two samples were analyzed on different genotyping platforms, we limited analysis to the intersection of SNPs between the two platforms. However, this complication did not cause significant concern, because the intersection contained 56,972 SNPs, a figure that is more than sufficient to produce reliable results for most analyses.

2.3.5 Principal components analysis for inference of population affinities

Principal components analysis (PCA) was performed using the program EIGENSTRAT [43, 44]. We ran PCA on the Qatar sample combined with all HGDP samples and plotted all the samples onto the resulting principal components. To investigate the possibility of admixture, we also constructed principal components on a subset of HGDP population samples and subsequently plotted the Qatar groups onto these principal components. Although there alternative interpretations for these PCA plots, one interpretation is that there was recent admixture within the focal population between the two ancestral populations whose points appear in clusters that flank the focal population when projected on the principal components [3]. Other interpretations include genetic drift between subpopulations, but this interpretation is only considered to be likely when the ancestral populations contribute the same proportions of ancestry to each subgroup [45].

2.3.6 Inference of pairwise IBD blocks

As a first pass inference of regions of the genome within each individual that are identical by descent (IBD), we applied PLINK to find these regions of homozygosity. The approach identifies spans of homozygosity within single individuals that may be consistent with considerable levels of autozygosity, and quantifies the range of inter-individual variation in this feature.

2.3.7 Correlations between genetic ancestry and surname lineage

Surnames of the individuals were sorted with knowledge of the local provenance of many of the family names into bins labeled Arab, African, Asian, and Persian as well as some pairwise ambiguities. Qatar has a small population with few, and usually common, surnames, but when a names origin was in doubt, we relied on the expertise of Qatar historians or at last resort, marked the surname as unclassified. These coded bins were then tallied by frequency in the three Qatari subgroups that had been identified by the STRUCTURE analysis. To assess the significance of the correlation between surnames and genetic ancestry, we created two binary distance matrices, one for surname origins and another for genetic subgroups, and submitted these matrices to the R package Mantel [46].

2.3.8 Patterns of decay of linkage disequilibrium

After using the intersection of filtered SNP sets for all population samples, we measured LD using the PLINK –r2 command to estimate the correlations between each marker pair genome-wide within each sample group. The correlation between SNPs as calculated by PLINK is a measure of the correlation between genotypes, as represented by minor allele counts, rather than haplotypes, as r^2 is usually portrayed. This change is purely for computational efficiency as calculations for haplotype correlations are significantly slower than for genotypes correlations and would become unwieldy genome-wide. However, these two values of r^2 do not differ significantly [39]. To increase efficiency further,

we limited the comparisons to SNPs less than 500 Mb apart. After binning these estimates by kilobase and averaging the estimates in each bin, we compared the calculated correlations between SNP pairs and the respective distance between the SNP pairs for all population samples.

2.4 Results

Analysis of the Qatar sample reveals three distinct subpopulations which differ in proportions of ancestral populations, degree of consanguinity, runs of homozygosity, and rate of LD decay. The ancestry of the three groups corresponds well with an Arabic group, an Asian group, and an admixed African group with other population genetic features resembling their respective ancestral populations. Furthermore, the origin of each subgroup correlates well with origin of the surnames of the individuals in each group.

2.4.1 Inference of population substructure within Qatar

Runs of the program STRUCTURE have been widely applied to yield an unsupervised clustering of individuals into affinity groups that appear to yield an approximation to a multi-locus panmictic collection of genotypes [40]. In a population that is suspected of having a high level of consanguinity, we need to proceed with caution. At first we attempted to use an extension of STRUCTURE that is designed specifically for an inbred population, but this turned out to be suitable only for highly inbred, partially selfing organisms, and results were not satisfactory [47]. When we used STRUCTURE, the results were quite

reasonable. As is often done, we ran the program with different prior guesses at the number of subgroups, including 2, 3, 4, and 5. The results are plotted as in previous STRUCTURE analysis and appear in Figure 2.1 [48, 49]. The k=3 and k=4 models fit best to the data with similar likelihood scores. Following parsimony and recommendations in the STRUCTURE software documentation, we took the k=3 run and separated individuals into three groups according to the STRUCTURE clustering.

We next turned to PCA as a means of identifying not only the affinities among these three groups of Qatari individuals, but their relations to other human population groups. For the latter we used the data from the HGDP, a collection of over 1000 individuals from 52 population groups spaced across the globe [41]. We first displayed the three Qatari subgroups in relation to the major human population groups (Figure 2.2). The three primary clusters of the Qatari are visually confirmed in the PCA plots. There is a very clear co-clustering of the Qatar1 group with the people of Middle Eastern origin. Qatar2 tends to show a greater affinity with Asian samples, although it is much more dispersed and partially overlaps with a few of the Qatar1 individuals. Qatar3 is clearly the most strongly African, and also has the greatest dispersion, much like the PCA plots of African Americans [50].

We further explored the relationships between the population samples by plotting smaller groupings of the HGDP populations and then displaying the Qatar samples with respect to the PCA loadings inferred from only the selected HGDP populations. This latter approach allowed us to investigate the unknown provenance of the Qatar samples with respect to the known HGDP samples as well as observe the extent of admixture, if any, in the Qatar samples. This analy-

sis showed again the tight clustering of the Middle East samples with the Qatar1 subgroup, and the spreading of both the Qatar2 and Qatar3 from this primary cluster (Figure 2.3A). The Qatar2 group stretches out slightly from the Middle Eastern populations to the Asian populations, while Qatar3 extends substantially toward African populations (Figure 2.3B). Figure 2.3C includes only the Qatar samples with Middle East and African samples of HGDP, and robustly shows the same result. Finally, limiting the set of HGDP Asian population samples to Chinese samples still preserves the affinity between Qatar2 and Asian samples, with the closest Asian group being the Uyghur. In sum, the impact of the trade along the axis from the Persian Gulf to the Indian Ocean is evident in the genetic make-up of present-day Qatar.

2.4.2 Consanguinity and runs of homozygosity

For each individual in the sample, we calculated Wrights inbreeding coefficient (f) from the allele frequencies in each population group and the homozygosity of the individual in question. The f coefficient calculated by PLINK is based on the number of observed and expected homozygous sites across the genome of each individual given the allele frequencies of each locus in the genome. Details on the calculation are given in the PLINK paper [39]. Calculating f permitted the distributions of the degree of inbreeding for each Qatari population subgroup to be assessed (Figure 2.4). Qatar1 shows a distribution akin to that seen in other Arab populations, with more than 10% of the sample having an inbreeding coefficient higher than that of offspring of first cousins (f = 0.125). But even though some individuals display significant consanguinity, there are nevertheless many individuals that appear to have no signature of inbreeding

at all. Qatar2 shows a much lower level of consanguinity. Surprisingly, Qatar3 has a marked tendency toward negative f values, consistent with a pattern of marriage following the trends of negative assortative mating [42]. The magnitude of the negative f-coefficients is surprising, and exceeds that seen in African Americans [51]. Applying the Wilcoxon signed rank test as implemented in the R statistical package, the f values for the Qatar3 group are significantly skewed toward values less than 0 (p-value = 7.63e-06).

A quantile-quantile plot indicates how the degree of consanguinity compares to the Bedouin sample of HGDP, a group known to practice frequent first-cousin marriages (Figure 2.5) [52]. There exists a good correspondence, apart from two individuals who appear more strongly consanguineous than any of the Bedouin samples, between Qatar1 and the Bedouins. Sample size is taken into account when creating quantile-quantile plots in that quantiles for smaller samples are interpolated to match those of larger datasets, so the outliers are probably not due to differences in sample size [53]. When examining runs of homozygosity, these two individuals have higher fractions of their genome contained in the runs when compared to the mean Qatari f, further supporting the idea that these individuals are highly consanguineous. However, there appears to be some trend toward less consanguinity for most individuals in Qatar2, even though this group also retains a few highly inbred individuals. Similar to what was seen in the previous histograms, the Qatar3 subgroup is remarkably non-consanguineous relative to the Bedouin sample.

PLINK is able to identify runs of homozygosity and, with a few assumptions, these runs of identity-by-state can be equated to runs of identity-by-descent. The implication is that the level of consanguinity may drive large portions of

the genome to have descended from a single common ancestor several generations in the past. Plots of the spans of homozygosity show that the Qatari sample has a wide range of homozygous blocks (Figure 2.7A), consistent with the variance in the degree of consanguinity (Figure 2.4). The contrast between the pattern of IBD sharing in the Qatari and the European-American samples (Figure 2.7B) is striking, especially in the variance between samples of each dataset. The European-American samples are much more even in the regions of IBD, lacking both tails of the distribution borne by the Qatari, which contain some samples with a relative surplus and others with a remarkable paucity of IBD regions.

2.4.3 Correlations between genetic ancestry and surname lineage

A Mantel test comparing Qatari subgroups and surname origins indicates highly significant (P-value = 0.0001) correlations across the 3 population groups in the frequency of these name classifications, with the Qatar1 having mostly Arab surnames, Qatar2 having a large Persian component, and the Qatar3 population appearing to be the most diverse and having the largest African component (Figure 2.6). In general, the genetics and this broad surname analysis appeared to be concordant.

2.4.4 Decay of linkage disequilibrium

Pairwise linkage disequilibrium among pairs of SNPs is a fairly sensitive indicator of the past history of recombination and genetic drift. When we tallied the pairwise r^2 for SNP pairs, binned them by distance in bp separating the SNPs along the genome (out to a maximum of 70 kb) and plotted the bin averages for each of the three Qatar subgroups, we see a strong difference among each group. In particular, the Qatar1 group, shows the slowest decay of LD, in keeping with its identity as largely Arab and consistent with its history of consanguinity. In fact, Qatar1 has a rate of LD decay even slower than the HGDP Bedouin sample (Figure 2.8A). Care was taken to do these comparisons with subsamples of the same sample size, as it is known that larger samples identify more recombinants and skew the LD downward.24 In the Qatar3 subgroup, the pattern of LD decay is similar to that seen in African samples of the HGDP (Figure 2.8B). LD is known to decay faster in Africa, most likely due to the larger and more long-term effective population size in Africa, and the fact that this population did not pass through an out-of-Africa bottleneck [54]. In sum, there is little surprise that the Qatar3, whose genetic affinity with the African populations had been identified by PCA, also shows a pattern of LD decay similar to Africans.

2.5 Discussion

The primary finding of the present report is that genetic variation among the current Qatari population is remarkably structured, and that this deep structure has been driven by historical migration and settlement in the area. We find that the Qatari can be largely divided into three primary affinity groups, one that

is of Arab origin and may be descendants of the Bedouin tribes; another that has strong affinity with Iranian (Persian) and other more eastern populations, including central Asia (such as the Uyghur); and a third which has a strong affinity with Bantu-speaking Africans. The latter two groups show strong patterns of admixture, with individuals showing a continuous spread of genetic affinity from the Middle Eastern toward the Asian and African populations respectively. The three groups demonstrate a strong correlation with family name supporting the local narrative on population history.

There is not a great wealth of literature on the genetic structure of the Qatari against which we can compare the present findings. A few studies have established some features of other Middle Eastern population samples, and the studies of the population of Saudi Arabia have advanced well. Previous studies examined the pattern of mtDNA variation in a Saudi sample, with a focus on testing whether the Saudi peninsula is peopled by remnants of the expansion out of Africa some 150,000 years ago [55]. The mtDNA lineages, because of their lack of recombination, retain clear information about maternal lineages, but because they do not recombine, they represent only one sampling of the myriad genealogical processes that occurred. The Saudi samples possessed both African lineages (20%) and eastern lineages (e.g. matching India and Central Asia) (18%), but the bulk was from a more northern origin (62%). This result suggests that, like the Qatari peninsula, the Saudi population harbors a diverse array of genetic contributions following centuries of active trade, and is not simply a relic of the ancient out-of-Africa migration. Patterns of Y chromosome variation are largely consistent with the mtDNA [56].

The pattern of historical influx and admixture in Qatar is strongly different

from patterns seen in Europe, where there is a remarkably clear pattern of isolation by distance [57, 2, 58]. Even India, which also has had much population movement and a strong impact of caste structure, retains a strong geographical component to its genetic structuring [59]. Historical patterns of migration and trade seem to dominate the pattern of influx of genetic variation into the Qatar Peninsula, and the drive to the trading centers and large expanses of desert result in an abolition of patterns of isolation by distance. In this context, our primary finding of three distinct groups appears to match well with Qatars migratory history.

The pattern of consanguinity, particularly the accepted practice of first-cousin marriages, has resulted in a high level of consanguinity, and as importantly, a huge inter-individual range of variation in Identity-by-Descent (IBD) sharing among the people of Qatar. The pattern of consanguinity is radically different among the three subgroups that we identified. These population-level findings have immediate and profound consequences for the practice of medical genetics in Qatar, and for the design and implementation of association testing in the future. The population is remarkably heterogeneous and structured. Ignoring this structure will lead to errors, both in individual diagnosis and in population-wide inference of SNPs that inflate risk of disease. It is also likely that these observations will be important in determining the genetic components involved in efficacy and adverse effects of pharmaceuticals in the different Qatar subpopulations. These studies need to be conducted in the context of the knowledge of which subgroup each individual has the strongest genetic affinity in order to draw accurate conclusions.

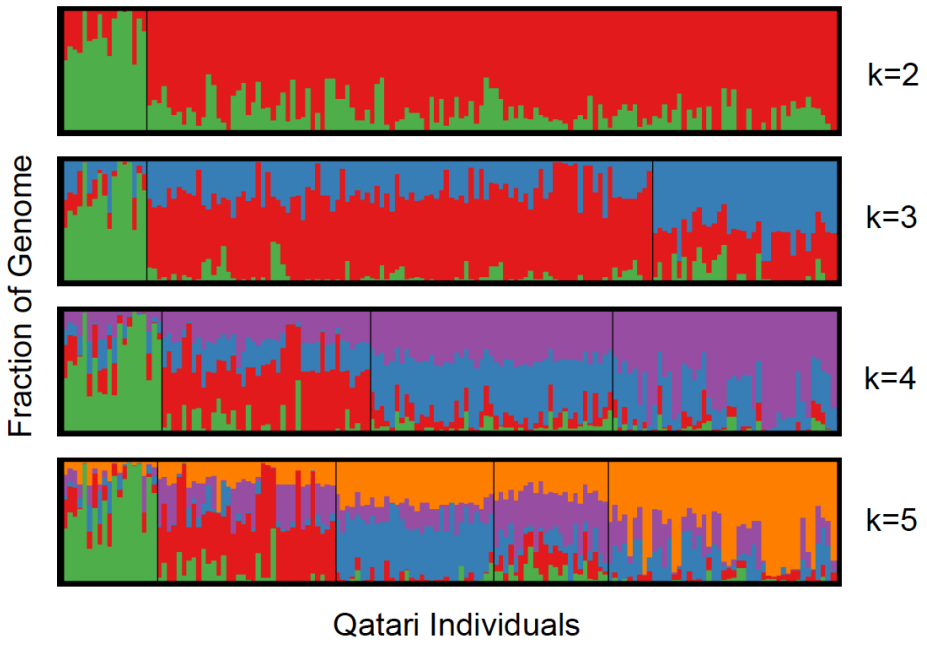


Figure 2.1: Qatari STRUCTURE results

Analysis of admixture using the program STRUCTURE assuming 2, 3, 4, and 5 subpopulations. The plot represents each individual as a thin vertical column. The proportion of each color in each column indicates the proportion of an individuals genome originating from one particular (but arbitrarily colored) subpopulation. For k=3, we arbitrarily label these subpopulations Qatar1 (red), Qatar2 (blue), and Qatar3 (green), and assign each individual to a subpopulation based on plurality.

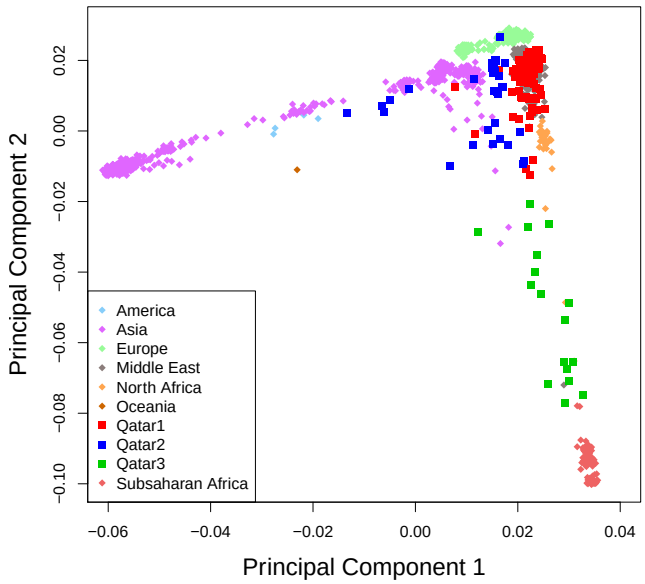


Figure 2.2: PCA of HGDP and Qatar samples

PCA plot of Qatar1, Qatar2, and Qatar3 (as defined by the Structure analysis in Figure 1) and population samples from the Human Genomic Diversity Project (HGDP). Qatar1 clusters well with other Middle Eastern samples. Qatar2 spreads away from the Middle Eastern cluster toward the Asian samples. Qatar3 spreads away from the Middle Eastern cluster toward the African samples . The interdigitation of the Qatar2 and Qatar3 samples could indicate recent admixture.

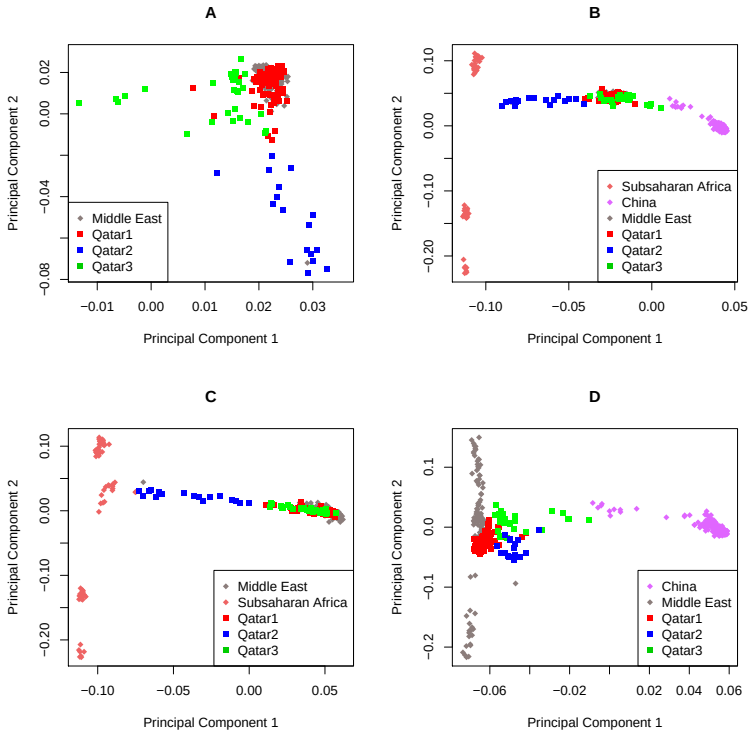


Figure 2.3: PCA plots revealing relations to the HGDP samples and the extent of Qatari subgroup admixture

See Appendix A for full caption.

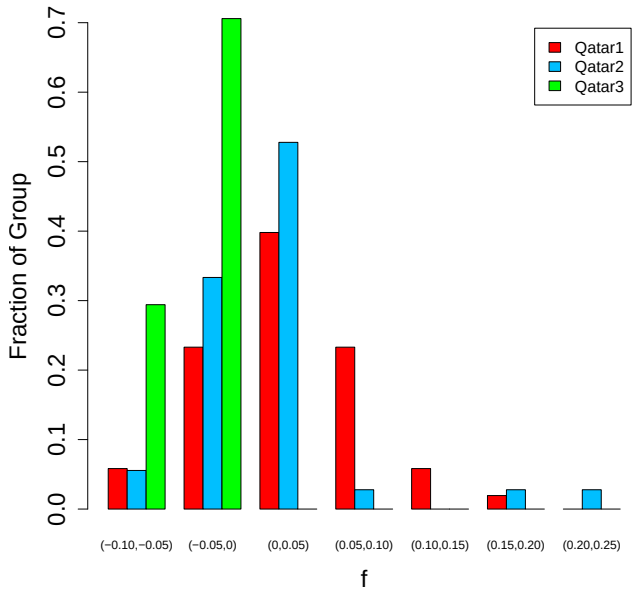


Figure 2.4: Distribution of the degree of consanguinity in each Qatar subgroup

The distributions of consanguinity are significantly different across the three Qatari subgroups. Qatar1 shows the highest degree of consanguinity while every individual in Qatar3 has an unusually low level of consanguinity. Two tests of the statistical significance of differences among these groups in consanguinity were performed: Kruskal-Wallis Test: $p<0.001$, and ANOVA: $p<0.001$.

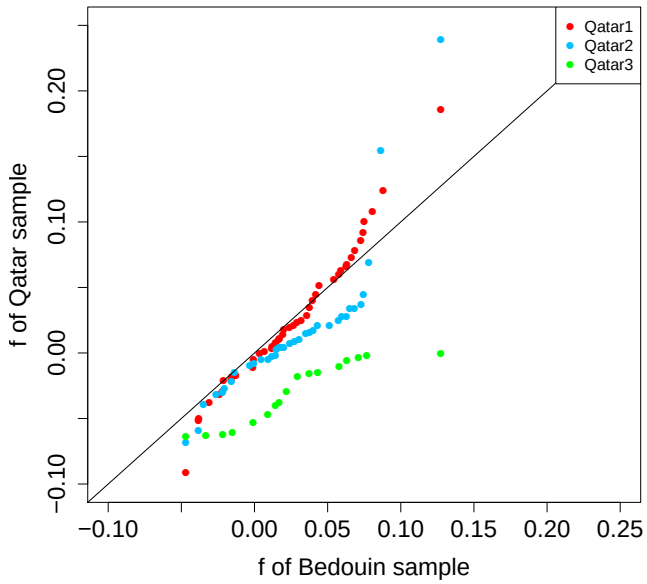


Figure 2.5: Analysis of the degree of consanguinity across the Qatari subgroups as compared to the HGDP Bedouin sample

Quantile-quantile plot comparing the Wrights inbreeding coefficient (f) as calculated with PLINK for each individual in each Qatar subgroup with the coefficients of each individual in the HGDP Bedouin sample. The plot indicates that the Qatar1 subgroup contains individuals with higher levels of consanguinity than individuals in the Bedouin sample. The Qatar2 subgroup contains individuals with a lesser degree of consanguinity (the trend of points below the diagonal) compared to individuals in the Bedouin sample, although there are two outlying individuals with unusually high consanguinity. Finally, the Qatar3 subgroup appears to be far less consanguineous than the Bedouin sample.

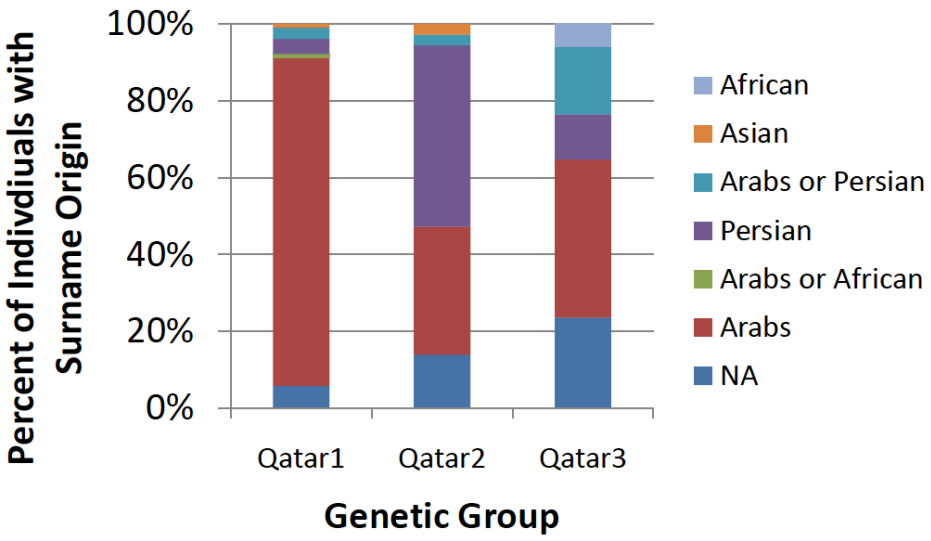


Figure 2.6: Qatari surnames and genetic classifications

Composition of surnames within each of the three Qatari groups is indicated by color coding according to surname frequency within those groups. The genetic classification of Qatar1, Qatar2, and Qatar3 are significantly correlated with surname origins (Mantel Test, $p<0.0001$).

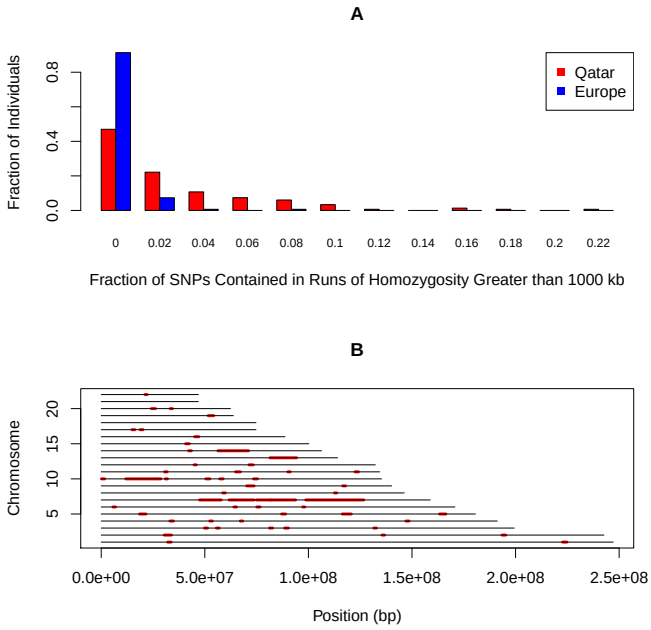


Figure 2.7: Spans of Qatari genomes that are homozygous

A. The fraction of each individual's genome that is contained within runs of homozygosity is plotted as a histogram for a sample of approximately 150 Qataris and 150 European-Americans. The minimum length of each is 1000 kb or 100 SNPs. The Qatari exhibit greater variance in homozygosity relative to the European-Americans. **B**. The runs of homozygosity for a single Qatari individual. The x-axis is the position along a chromosome (0 - 250 Mbp) and the y-axis is the chromosome number. Each colored segment represents a block of sequence in which SNP marker genotypes are homozygous.

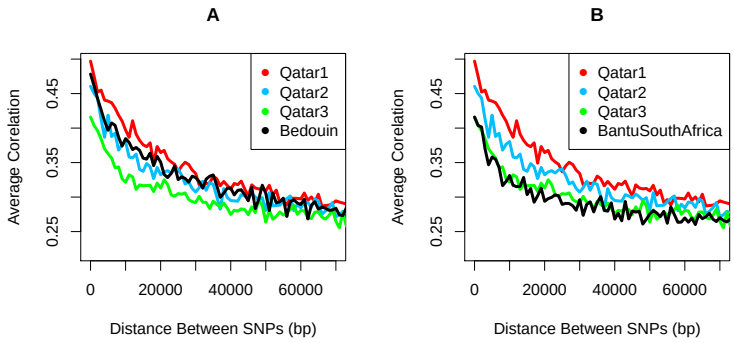


Figure 2.8: Linkage disequilibrium decay across the genomes of the Qatari subgroups and two HGDP population samples

A. Linkage disequilibrium (LD) for pairs of SNPs less than 70 kb apart were calculated as the squared correlation coefficient (r^2). Calculations were done on a standard sample size (n = 5) of randomly selected individuals in each Qatar group. SNP pairs were partitioned into bins by each kilobase (kb) distance, and mean bin r^2 was plotted. The Qatar1 group has the highest LD consistent with their higher degree of consanguinity. Qatar2 is intermediate, and the Qatar3 group has the lowest LD between SNPs, consistent with a large African component in their genome. The Bedouin HGDP population sample appears to fall between that of the Qatar1 and Qatar2 groups. **B**. The decay of LD of the three Qatari samples is re-plotted here along with the Bantu South African sample of the HGDP set. The LD decay of the Bantu South African population sample overlaps with that of Qatar3, consistent with the Qatar3 sample being largely of African origin.

CHAPTER 3

ABERRANT TIME TO MOST RECENT COMMON ANCESTOR AS A SIGNATURE OF NATURAL SELECTION

Haley Hunter-Zinck and Andrew G. Clark

3.1 Abstract

Natural selection inference methods often target one mode of selection of a particular age and strength. However, detecting multiple modes simultaneously, or with atypical representations, would be advantageous for understanding a population's evolutionary history. We have developed an anomaly detection algorithm using distributions of pairwise time to most recent common ancestor (TMRCA) to simultaneously detect multiple modes of natural selection in whole-genome sequences. Since natural selection distorts local genealogies in distinct ways, the method uses pairwise TMRCA distributions, which approximate genealogies at a non-recombining locus, to detect distortions without targeting a specific mode of selection. We evaluate the performance of our method, TSel, for both positive and balancing selection over different time-scales and selection strengths and compare TSel's performance to that of other methods. We then apply TSel to the Complete Genomics diversity panel and recover loci previously inferred to be under positive or balancing selection.

3.2 Introduction

Natural selection is the driving force behind adaptive evolution. The ability to detect regions of the genome that have undergone natural selection has increased our understanding of function and evolutionary history of many loci [14]. However, natural selection takes different forms, all of which are informative, and current selection inference methods each target only a subset of natural selection scenarios [27]. Given the importance of many modes and degrees of natural selection, a method that could detect multiple, atypical, or combinations of selection scenarios simultaneously would be both convenient and advantageous. Furthermore, given the current amount and continuing accumulation of sequencing data, designing methods for whole-genome sequences, rather than genotype data, will harness additional genetic information.

Detecting anomalous genomic sites has long been the foundation of natural selection tests [28]. This approach is susceptible to both false positives and false negatives, but using a statistic that better distinguishes selected and neutral sites, using multiple statistics, or both could improve performance. To be truly general, an anomaly detection algorithm should also make use of any number of features and account for correlations among features. Furthermore, instead of using statistics based directly on extended haplotypes or sequence diversity, which are characteristic signatures of particular modes of selection, a general natural selection algorithm should use a universal measure that responds uniquely to each mode of natural selection. Theory and empirical studies demonstrate that natural selection distorts local genealogies in distinct and systematic ways, and exploiting these distortions could lead to a more general method [60]. Although inferring local ancestral recombination graphs genome-

wide is still computationally prohibitive, methods for inferring pairwise time to most recent common ancestor (TMRCA) distributions, which are approximations of local genealogies, are now available [5]. As we will show, key advantages of detecting selection based on TMRCA include detecting multiple and uncharacterized forms of selection and making full use of whole-genome sequence data.

TMRCA is closely related to identity-by-descent (IBD); a threshold on TMRCA in a local genealogy defines local IBD between two individuals in an unrelated population sample [16]. Several studies have used IBD to infer the presence of positive selection in the genome [10, 17]. However, using TMRCA directly, instead of approximating the continuous metric with a binary call such as IBD, could prove both more effective and more versatile in application. Although the estimation of recent TMRCA with current methods has high variance, IBD segment length has even greater variance due to the fact that it depends on the closest pair of recombination events, limiting its utility for inference [5, 16].

Here, we develop an anomaly detection test using TMRCA that can simultaneously detect multiple modes of selection. There are many advantages to our formulation of the selection inference problem. Anomaly detection resembles the intuition behind many selection tests, and our implementation can include any number and type of features and account for correlations between these features. By using the input data itself to construct a model of neutrality, we also account for demography without specifying an external demographic model. Our method also has the capability to detect selection in sequence data in a scalable fashion. Furthermore, using features derived from pairwise TMRCA dis-

tributions exploits our knowledge about how natural selection causes local and systematic distortions in genealogies and creates a general test that can detect uncharacterized, atypical, or combinations of modes of natural selection acting on a single locus. We discuss the performance of the method, which we call TSel for TMRCA Selection, in simulated data, compare the method's performance to other selection inference methods based on simulated data, and then apply our method to the Complete Genomics diversity panel [61].

3.3 Methods

3.3.1 Features of exact pairwise TMRCA distributions

We extracted the exact pairwise TMRCA values for each pair of chromosomes from simulated coalescent trees output for each non-recombining locus. Simulations are described below. For clarity, we will refer to the pairwise TMRCA values extracted directly from the simulated coalescent trees as the exact pairwise TMRCA values to distinguish them from inferred pairwise TMRCA values analyzed later in the study. From the distribution of exact pairwise TMRCA values at each non-recombining locus, we calculated a variety of features, including the average, maximum, median, variance, skewness, kurtosis, a bimodality coefficient, fraction of pairs equal to the maximum, and various quartile values. We also normalized each replicate's exact pairwise TMRCA distribution to be between 0 and 1 and calculated relevant features on these normalized distributions as well. Because using irrelevant feature may decrease performance, we calculated the Laplacian Score on each of the features to select the most discrim-

inative features of the set [62]. The Laplacian Score is an unsupervised feature selection method that compares each feature to the global similarity of all the samples to select features that are most discriminative between cluster in the data. The Laplacian Score greatly outperforms feature selection that uses only the variance as a ranking metric. We select the top 95% of the features to include in the classifier, and each non-recombining locus was then represented by a vector of the extracted features.

3.3.2 Anomaly detection algorithm

In the TSel method, we applied a simple anomaly detection algorithm to the features of exact pairwise TMRCA distributions. This algorithm uses the Mahalanobis distance in which the mean and covariance matrix are calculated on a set of putatively neutral data samples [63]. Before calculating the Mahalanobis distance, we removed invariant and correlated features, and selected the most discriminative features using the Laplacian Score. Because the Mahalanobis distance assumes normally distributed features, we then transformed the features using a Box-Cox transformation with the help of the R package geoR to ensure normality [64]. Using the transformed features, we calculated the mean and covariance for these features over all neutral loci and then the Mahalanobis distance for each sampled locus.

TSel is implemented as the R package tsel and is available on the Comprehensive R Archive Network (CRAN).

3.3.3 Simulations

We generated simulated data using the program MSMS (version 3.2rc Build:147), sampling 100 chromosomes for a locus size of 10 Mb with a constant recombination rate of 1.0×10^{-8} and a mutation rate of 1.1×10^{-8} [65, 66]. For computational reasons, we restricted recombination such that recombination events can occur only every 100 bp. This restriction did not appear to affect simulation results as the mean non-recombining window size was well above the 100 bp minimum. The simulator also output coalescent trees for each non-recombining window and diversity statistics π, Watterson's θ, and Tajima's D over 10 kb windows.

In order to demonstrate the method's performance on different modes of selection, we simulated loci undergoing complete hard sweeps, partial hard sweeps, complete soft sweeps, and overdominance. We also varied the time of equilibrium and the strength of selection for each scenario. Hard sweeps began from one copy of the selected allele and the time of sweep completion was set to 40, 400, and 4,000 generations in the past. We used an additive model for selection coefficients, and the selection strength for individuals homozygous for the selected the allele was set to 0.1, 0.01, and 0.001. For partial hard sweeps, we set the final frequency of the selected allele to 0.75, and for soft sweeps, we set the initial allele frequency of the selected allele to 1%, to simulated selection from standing variation. To assess performance in different demographic scenarios, we simulated data with an effective population size of 1,000 and 10,000 individuals.

For overdominance, we parameterized selection by the approximate time in the past at which equilibrium was reached and set this value to 400, 4,000, and

40,000 generations in the past. We estimated this time from the simulated data and fed the initial time of selection, the generation value plus the time to equilibrium, to the simulator. Selection began from one copy of the selected allele. We set the selection coefficient for those individuals heterozygous for the selected allele to 0.1, 0.01, and 0.001 and homozygous individuals had a selection coefficient of 0. To ensure the allele was not lost, we conditioned on the presence of the allele in the forward simulations. Because alternative balancing selection tests require information from an outgroup relative to the tested sample, we simulated the selection scenarios with an outgroup diverging 6.5 million years ago to approximate human and chimp divergence [67].

3.3.4 TSel performance

To evaluate Tsel's performance, we generated 10,000 simulated replicates of each neutral scenario and 1,000 replicates of each selection scenario. We extracted the exact pairwise TMRCA features at the center of the simulated locus, the location of the selected variant if present, for each replicate and ran the TSel algorithm using the neutral data alone to select features, transform features, and then calculate the mean and covariance matrix. We then calculated the Mahalanobis distance on both the neutral and the selected replicates and compared performance using ROC curves [68].

For comparison we also assessed the performance of other methods on the simulated data. For hard and soft sweeps, the positive selection scenarios, we compared TSel's performance to the iHS and excess IBD methods [69, 10, 17]. The iHS test is frequently used, unbiased by demography, and effective as a

test for recent positive selection. We did not use the CMS test because this test requires multiple cross-population statistics and we wanted to compare TSel's performance to that of other methods that can detect selection within a homogeneous sample [70]. Because of computational time, we calculated the iHS statistic on only 1,000 neutral and 100 selection simulated replicates using the R package *rehh* and extracted the median absolute value of the iHS score for a 100 kb window around the selected locus [71].

Several studies have successfully utilized excess IBD sharing to detect positive selection [10, 17]. We chose to compare TSel's performance with excess IBD because of its relationship with TMRCA. In population samples IBD is defined by drawing a threshold across a genealogy at a particular time in the past [16]. Although many inference methods do not explicitly model IBD in this manner, inference power as a function of IBD block length implicitly assumes this definition. We determined IBD status in our simulations by drawing a threshold 100 generations in the past, roughly the limit of inference power with current methods, and calling chromosomes that coalesced more recently than this threshold IBD. We then calculated the fraction of pairs at the selected locus that were IBD and constructed ROC curves using these values. We note that our approach was slightly different than the previously employed approaches, which used the posterior probabilities of IBD rather than making discrete calls. But since we obtained the local genealogies from simulated data, our definition of IBD does not need to be probabilistic as it provides the exact IBD calls.

To compare TSel's performance on balancing selection, we ran the Hudson-Kreitman-Aguadé (HKA) test [72]. The HKA test is a standard test for balancing selection in genetic data. Because of increased running time, we only ran the test

on 1,000 replicates for the neutral scenario and 100 replicates for each overdominance scenario. We ran Jody Hey's implementation of the HKA test (available at `http://astro.temple.edu/~tuf29449/software/software.htm`) using a window size of 10 kb, two loci, and one sample from the outgroup, following the original test procedure. We then assessed the method's performance and compared power to TSel with ROC curves.

3.3.5 TSel performance with alternate features

The anomaly detection method is not limited to using features of exact pairwise TMRCA distributions, and other groups of features may also perform well. For comparison, we ran the method with features derived from diversity. We output π, Watterson's θ, and Tajima's D directly from MSMS for the same simulation scenarios that we tested with the exact pairwise TMRCA features but calculated over a 10 kb window. Again, we extracted the selected locus from each replicate and assessed performance with ROC curves.

We also analyzed performance with inferred pairwise TMRCA values instead of exact pairwise TMRCA values output by the simulator. We ran this check to ensure that TSel maintains improved performance with current TMRCA inference methods, and is therefore suitable for real data applications. We also compared performance with features derived from diversity to ensure that inferred TMRCA is not simply a proxy for diversity. We tested the features on complete hard sweep simulations, as described above, with an effective population size of 10,000. Instead of inferring pairwise TMRCA on all pairs of chromosomes, we ran PSMC on 50 pairs of chromosomes from our sample of 100 to

resemble within individual PSMC runs on real data. Because of PSMC runtimes, we simulated only 1,000 neutral replicates and 100 replicates for each complete hard sweep scenario. We then ran PSMC on the 50 chromosome pairs for each replicate and extracted the same features from the inferred pairwise TMRCA distributions as for the exact TMRCA distributions. After extracting these features, we ran TSel and compared TSel's performance via ROC curves for exact pairwise TMRCA features, inferred pairwise TMRCA features, and diversity features.

3.3.6 TSel performance when including selected sites

The anomaly detection method assumes that selected sites are rare in the data upon which we calculate the mean and covariance matrix. However, a portion of real data will be under selection, and it is important to assess the performance of the method when these data points are included. We used a complete hard sweep scenario with recent strong selection and an effective population size of 10,000 in order to test the effect of including selected sites. We included a range from 1% to 10% of data simulated under the selected scenario, calculated the mean and covariance on these data sets, and then constructed ROC curves. We then calculated the area under the curve (AUC) to assess the effect on performance for each percentage of selected data.

3.3.7 Application to Complete Genomics diversity panel

To exemplify our algorithm on real data, we used the Complete Genomics (CG) diversity panel consisting of 46 individuals from 9 different populations. CG generated the data with the Complete Genomics Analysis Pipeline version 2.0.0 [61]. The 46 individuals were sequenced to high coverage, approximately 80x average genome-wide, making these samples ideal for inference of pairwise TMRCA using the PSMC method.

Before running TSel on the CG diversity panel, we filtered extensively to avoid confounding factors. Li and Durbin note in their supplementary material that false positive variants increase the inferred TMRCA in all time intervals [5]. False negatives will change the scaling of the inferred values but may be easily accounted for by appropriately scaling the neutral mutation rate. To limit false positives due to sequencing or mapping errors, we marked variants as missing if the variants did not pass the CG quality thresholds, were indels, were within 10 bp of indels, or had more than twice or less than half of the average individual coverage depth. We also identified regions that had abnormally high TSel scores, probably due to mapping errors, such as within large segmental duplications, and excluded these regions from the analysis.

After masking variants and regions based on the above criteria, we ran PSMC on the chromosome pairs within the 46 individuals genome-wide. We then calculated the TMRCA distribution features listed above from the inferred pairwise TMRCA values for each 100 bp window. After running TSel, we consolidated the TSel scores for each 100 bp window into 10 kb windows by taking the median score. In order to avoid spurious hits, we discarded 10 kb windows that had more than 50% of the 100 bp windows missing and used the remaining

10 kb window values for subsequent analyses.

We submitted the top 1% of 10 kb windows to the program GREAT to examine gene ontology for the top TSel hits [73]. We also compared the overlap of our top 1% regions to the results of the recent positive selection scan on the 1000 Genomes Project data and a balancing selection scan of human data to ensure that TSel replicates regions previously inferred to be under positive or balancing selection [11, 12]. We further compared the top 1% regions to regions having an extremely deep TMRCA as inferred via the program *ARGweaver* [25]. We used 15 of the top 20 TMRCA estimates, excluding the regions with CNVs as these regions are filtered out in the TSel analysis. This comparison is especially interesting as it compares the results of using a pairwise versus multiple haplotype TMRCA inference approach.

3.4 Results

3.4.1 TSel performance

TSel exhibits excellent performance on hard sweeps, especially with a higher effective population size and stronger and more recent selection. ROC curves for TSel performance on complete hard sweeps for an effective population size of 10,000 are shown in Figure 3.1. TSel has an area under the curve (AUC) of 1.00 for stronger, complete hard sweeps and still shows some ability to detect weaker sweeps. For an effective population size of 1,000, TSel performance is random for ancient and weaker sweeps but still exhibits an AUC of 1.00 for the most recent and strongest sweeps (Fig. 3.2). Performance is lower for the

iHS method and the excess of IBD. iHS performance suffers slightly on complete hard sweeps because the iHS ratio becomes undefined for fixed alleles, but iHS still obtains an AUC of 0.87 for the strongest and most recent sweep with an effective population size of 10,000. The method using excess IBD shows no power to detect older sweeps, as expected since we defined IBD as coalescing more recently than 100 generations in the past, but still exhibits an AUC of approximately 0.5 for selection scenarios that completed even as recently as 40 generations from the time of sampling. TSel substantially outperforms iHS and excess IBD in detecting complete hard sweeps.

We also applied TSel to partial hard sweeps. TSel's performance is shown for an effective population size of 10,000 and partial hard sweeps ending with the selected allele at 75% frequency in Figure 3.3. Similar to the method's performance on complete hard sweeps, TSel exhibits an AUC of 1.00 for more recent and stronger partial hard sweeps, but different from the complete hard sweeps, iHS now shows similar performance to TSel. TSel's performance and that of iHS then declines as sweeps reach equilibrium in more ancient times or sweep strength decreases, but TSel maintains a higher AUC than iHS in these scenarios. For example, on the strongest but most ancient partial hard sweeps, iHS obtains an AUC of 0.69 while TSel reaches 0.86 . Excess IBD shows some power to detect the most recent and strongest partial hard sweeps, reaching a maximum AUC of 0.85, but still performs nearly randomly for other scenarios. As shown in Figure 3.4, results are similar for an effective population size of 1,000 but with reduced performance for all methods.

TSel performance suffers slightly on soft sweeps compared to that of hard sweeps, but the method still demonstrates power to detect sweeps from stand-

ing variation. Results are shown in Figure 3.5 and Figure 3.6. For the strongest and most recent soft sweep with an effective population size of 10,000, TSel outperforms iHS and IBD, obtaining an AUC of 0.94 compared to 0.38 and 0.51 for iHS and IBD respectively. Interestingly, TSel and iHS performance improves, to an AUC of 0.96 and 0.81 respectively, for moderate selection in the same time period, but TSel still outperforms iHS by a wide margin.

Having analyzed TSel's performance for positive selection, we now turned to examine the method's performance for balancing selection. TSel obtains a maximum AUC of 0.97 to detect more recent overdominance and a maximum AUC of 0.92 to detect ancient overdominance as shown in Figure 3.7 and Figure 3.8 respectively. The HKA test shows limited power to detect overdominance except for ancient selection in an effective population size of 1,000, where it reaches an AUC of 0.83 but is still outperformed by TSel.

It is important to note that TSel performance is dependent on the scaled time to selection equilibrium, not the actual generation time. Because TSel uses parameters derived from the distribution of scaled TMRCA values, performance will be best when the scaled time of selection equilibrium differs greatly from scaled expectation of the TMRCA in neutral trees. For example, ancient overdominance will create a much deeper tree than is expected under neutrality, while recent overdominance will create a separate mode of pairwise TMRCA values before the expected neutral TMRCA of the tree. This fact explains why different population sizes affect the performance over similar generation times. Overdominance scenarios reaching equilibrium at 4,000 generations in a population with an effective population size of 1,000 are equivalent in scaled time to the scenarios reaching equilibrium 40,000 generations ago in an effective pop-

ulation size of 10,000. The same reasoning holds true for the positive selection scenarios as well across different effective population sizes.

3.4.2 TSel performance with alternate features

To test if other feature subsets have equivalent performance, we examined TSel performance with exact pairwise TMRCA features, inferred pairwise TMRCA features, and diversity features (Fig. 3.9). Performance for TSel with TMRCA features versus diversity derived features is correlated, but TSel with exact or inferred TMRCA features outperforms that with diversity features. For example, in recent sweeps with an intermediate strength of selection TSel with exact TMRCA, inferred TMRCA, and diversity features reaches an AUC of 1.00, 0.98, and 0.94 respectively. Performance with inferred TMRCA features is lower than that with exact pairwise TMRCA most probably due to a variety of factors including inference errors and reduced number of pairwise TMRCA values. Performance of inferred TMRCA distributions will likely improve when all pairwise TMRCA values are included and inference methods improve. The slight difference in performance between inferred TMRCA features and diversity features is most likely a result of greater stochasticity in mutations compared to local genealogies. Diversity features are also derived on larger windows in order to include sufficient variation for calculation, effectively blurring local effects. The results demonstrate that inferred TMRCA is a distinct and more informative metric than measures of diversity for the inference of natural selection.

3.4.3 TSel performance when including selected sites

An important assumption of the anomaly detection method is that selected sites are relatively rare within the set of data upon which we calculate the feature means and covariance matrix. Since real data contains loci under selection, we tested the performance of TSel with varying fractions of selected sites included in the initial dataset. Results are shown in Figure 3.10. Although performance, as measured by the AUC, declines when we include more selected sites, the AUC is still 0.84 even when 5% of the data is under strong selection. Therefore, the method still has power to distinguish neutral and selected loci even when the dataset contains a substantial proportion of sites under strong selection.

3.4.4 Application to Complete Genomics diversity panel

Analysis of the top 1% of TSel hits with the program GREAT reveals 5 enriched biological properties including antigen processing and presentation of peptide or polysaccharide antigen via MHC class II, mammary gland specification, eyelid development in camera-type eye, columnar/cuboidal epithelial cell differentiation, and mammary gland formation. The top 1% of hits overlaps 6 regions from the CMS positive selection scan of the 1000 Genomes Project data and 3 of the inferred regions for the balancing selection scan of Leffler, *et al*. [11, 12]. Two of the replicated regions, one for the positive selection scan and the other for the balancing selection scan, are shown in Figures 3.11 and 3.12 respectively. Finally, we compared the top 1% of TSel hits to windows with the most ancient TMRCA identified via *ARGweaver* [25]. TSel hits overlap 4 of 15 top regions from *ARGweaver* after excluding the top regions with possible CNVs. Such an

overlap is encouraging replication for the pairwise TMRCA inference approach utilized in TSel.

3.5 Discussion

We have developed a powerful and flexible method that exhibits higher performance than current natural selection inference methods in a wide parameter space of simulated data. Furthermore, in real data, we have replicated loci previously found to be under both positive and balancing selection with a single method. TSel is more general than previous methods because the method detects any mode of natural selection that leaves a detectable distortion in local genealogies. These modes could include not only the classical mechanisms of natural selection but also combinations of selection modes or atypical presentations of known modes. Furthermore, the method accounts for demography by comparing loci to the data itself without specifying an external demographic model. Given the number of recent studies and range of demographic models for the European human population alone, this fact increases the ease of the method's application [6, 7, 9]. Lastly, because the inference of TMRCA requires whole-genome sequences, TSel takes full advantage of the growing accumulation of sequence data. These factors make TSel a powerful and flexible method for application to many datasets.

Using pairwise TMRCA is also an important development from pairwise IBD methods when analyzing population samples of unrelated individuals. Although IBD is readily defined within a pedigree, defining IBD in a population is harder to define consistently. Currently, IBD in a population is defined explicitly

by drawing a threshold backward in time across the coalescent tree or implicitly by power thresholds within current inference methods [16]. Furthermore, defining relatedness between individuals in terms of IBD reduces a vector of continuous statistics into a binary call, collapsing valuable information. Now that methods that infer TMRCA in sequence data are available, we have the ability to utilize a continuous metric of relatedness within a population. This approach could lead to better performance not only in selection inference, as shown above, but also in other IBD applications.

The method described here is similar in spirit to the composite of multiple signals (CMS) test but has important differences [70]. Most importantly, TSel uses features of pairwise TMRCA distributions, which approximate the local genealogies that are distorted in systematic ways by natural selection, and our method provides a simple framework to detect these distortions. Moreover, our method uses the Mahalanobis distance instead of using empirical distributions to model the data, which allows us to account for correlated features. Using information from feature correlations, the method detects not only rare feature values but also rare combinations of feature values. We did not compare the CMS test directly to the TSel method in the simulation studies because CMS relies on cross-population statistics which TSel's performance does not require. However, we did compare TSel to the iHS method, one of the statistics incorporated in CMS, in simulated data and also compared results of the TSel CG diversity panel analysis to CMS results on the 1000 genomes data, revealing 6 overlapping region.

However, several challenges remain. TSel is based on a statistic that ranks loci according to the Mahalanobis distance, but the method does not give a

threshold for determining significant deviation from neutrality. By taking the top 1% of loci, we hoped to examine the most extreme signatures of selection. That said, we stress here that even though it is tempting to define loci in discrete classes, natural selection in reality operates along a continuum of strengths, times and modes and that the TSel score recapitulates this continuum. Especially considering recent ENCODE results that suggest that 80% of the genome is functional, modeling loci not as neutral or under one mode of selection, but as a certain distance from neutrality has an intuitive appeal [74]. Another remaining challenge is to describe the mode of selection acting on each locus. A user could examine particular regions of interest by assessing the feature values of the locus along with allele frequencies in the region, and subsequent investigation into the functional annotation of the region could also help reveal which selective forces are at work. Lastly, although we have tested TSel performance in simulated data in different constant population sizes, we have not fully tested TSel on a complex demographic scenarios. Performance may suffer on particular modes of selection, especially sweeps, when confronted with populations that have undergone bottlenecks, while other types of selection, such as overdominance, may be easier to detect. But although we do not test the performance in complex demographic scenarios, the fact that TSel replicates loci previously inferred to be under both positive and balancing selection when run on the CG diversity panel, a sample with a complex demographic history, is encouraging for TSel's performance when confronted with complex demography.

To further address these challenges, it is worthy of note that TSel is not limited to using features of pairwise TMRCA. Any method statistic, along with features derived from diversity, cross-population statistics, or functional annotation, is easily incorporated into the method. Additional statistics would likely

improve performance or tailor TSel to detecting modes of selection of particular interest to the user. Future features of particular interest are those derived from complete genealogical trees. New methods are being developed to extend PSMC to incorporate multiple haplotypes that can not only increase the accuracy of recent TMRCA estimates but can also approximate or reconstruct local genealogies in full[75, 25]. With scalable methods to infer local genealogies we will be able to employ features such as tree length and height as well as tree imbalance to more accurately detect systematic distortions caused by natural selection in genetic data [76]. More comprehensive statistics in addition to higher coverage sequence data and larger sample sizes have the potential to elucidate more loci under selection and increase our understanding of the evolutionary history of any sample under study.

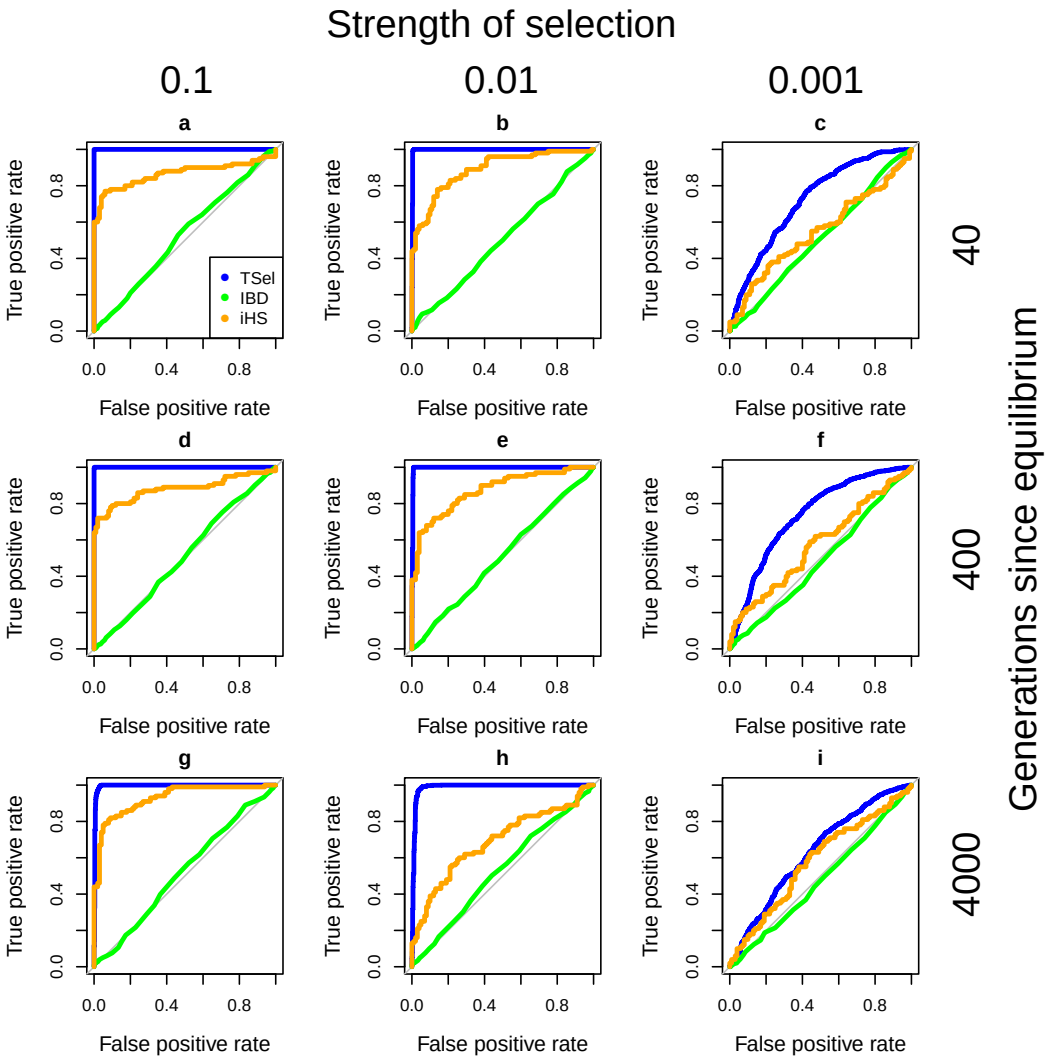


Figure 3.1: TSel performance on complete hard sweeps with an effective population size of 10,000

Performance is demonstrated via ROC curves. The x-axis of the grid corresponds to the strength of selection and the y-axis corresponds to the time of sweep completion.

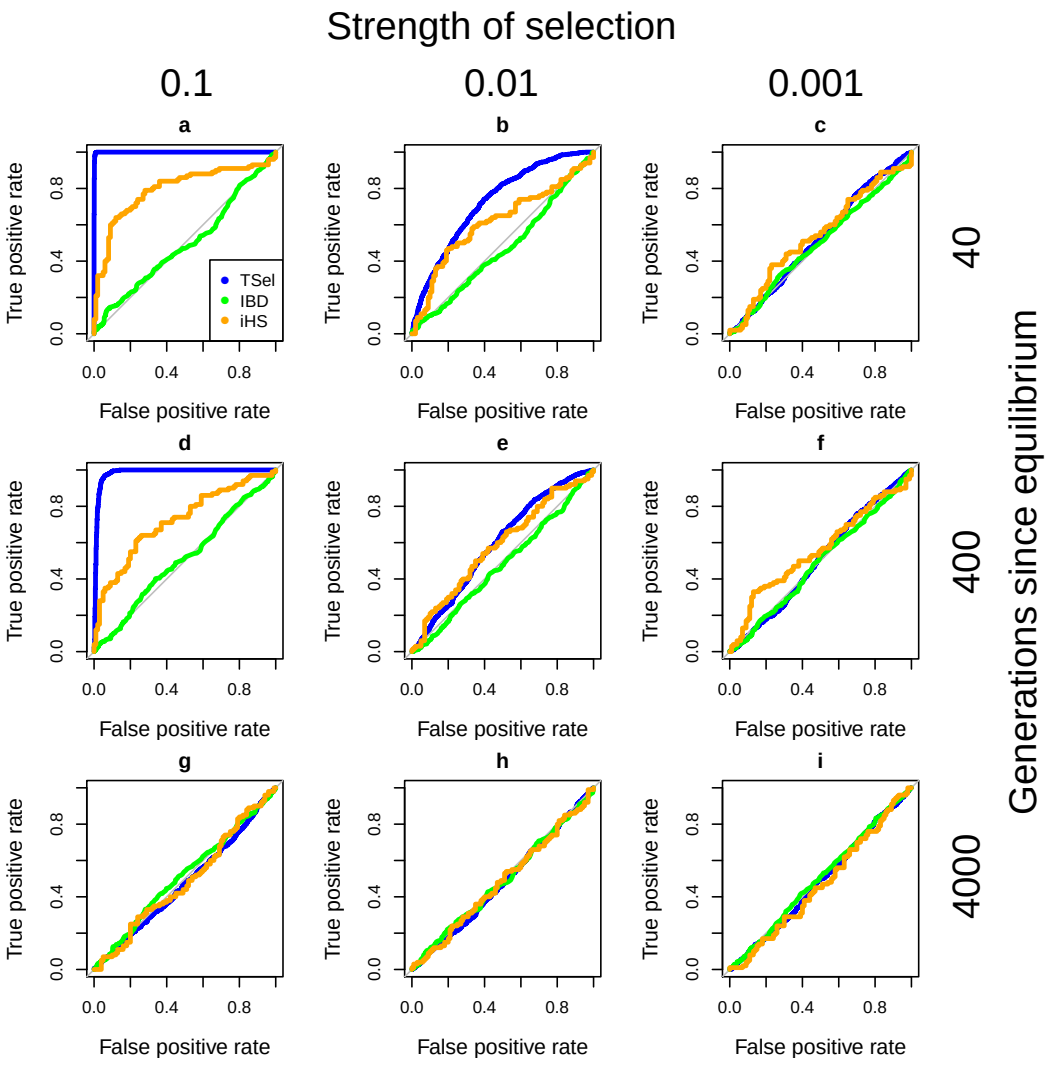


Figure 3.2: TSel performance on complete hard sweeps with an effective population size of 1,000

Performance is demonstrated via ROC curves. The x-axis of the grid corresponds to the strength of selection and the y-axis corresponds to the time of sweep completion.

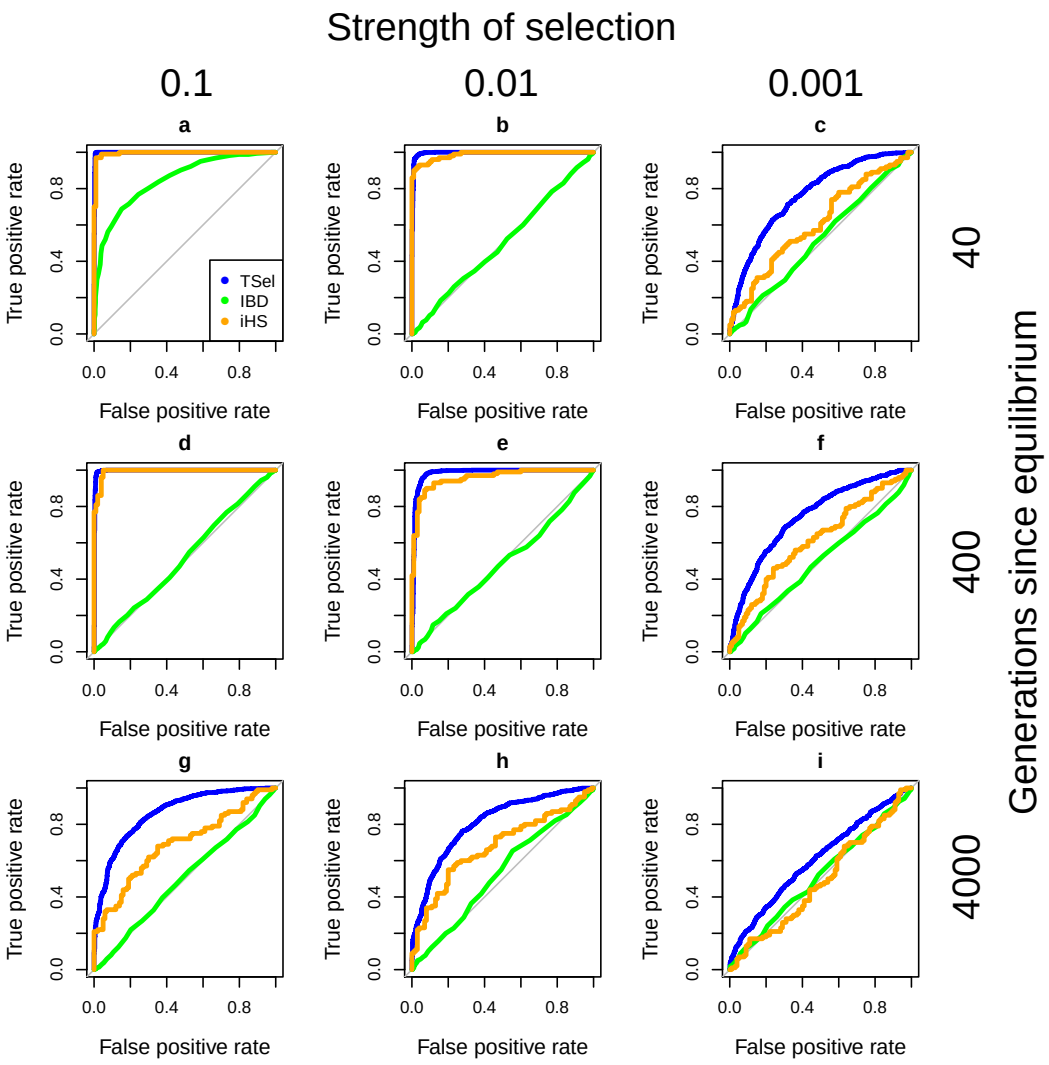


Figure 3.3: TSel performance on partial hard sweeps with an effective population size of 10,000

Performance is demonstrated via ROC curves. The final selected allele frequency of the partial hard sweep was set to 75%. The x-axis of the grid corresponds to the strength of selection and the y-axis corresponds to the time of sweep completion.

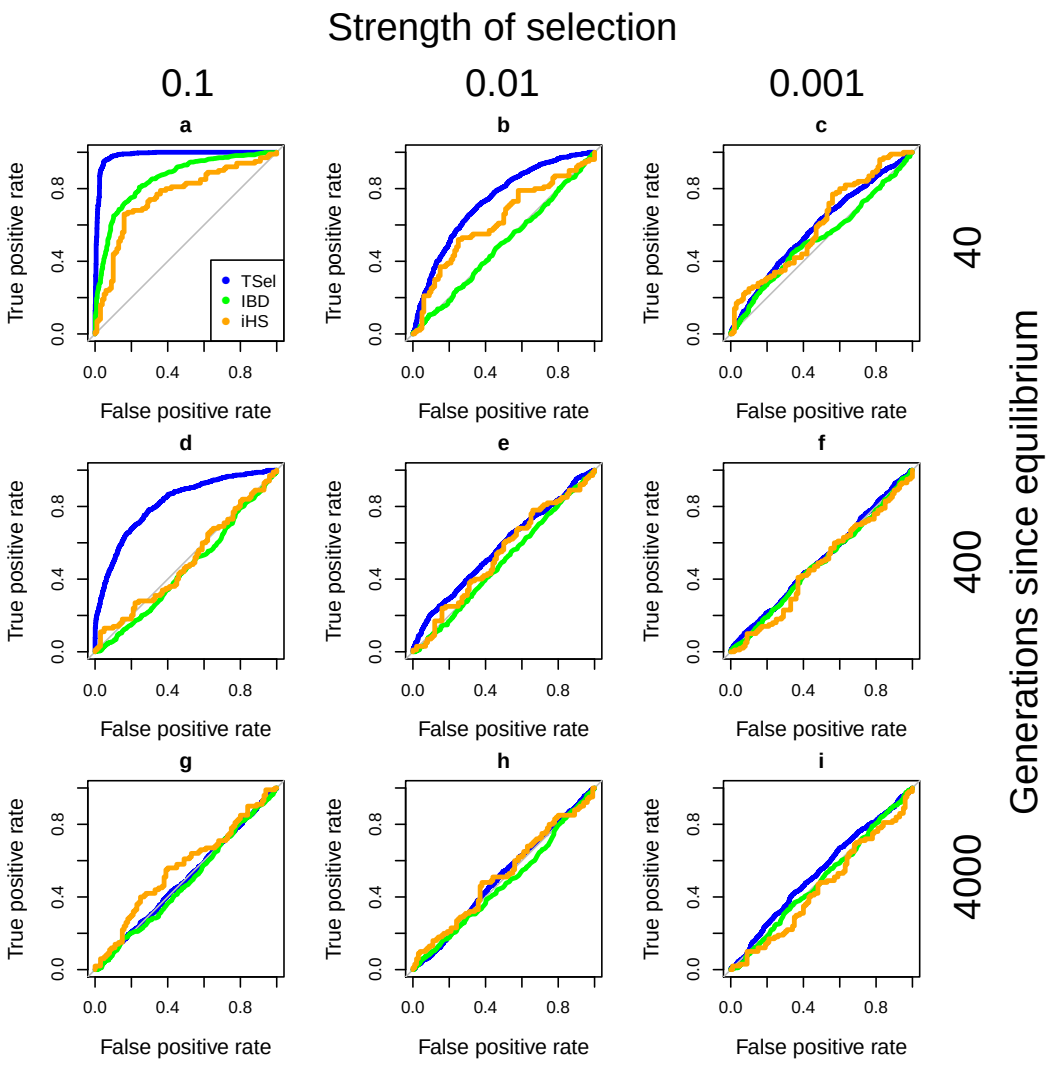


Figure 3.4: TSel performance on partial hard sweeps with an effective population size of 1,000

Performance is demonstrated via ROC curves. The final selected allele frequency of the partial hard sweep was set to 75%. The x-axis of the grid corresponds to the strength of selection and the y-axis corresponds to the time of sweep completion.

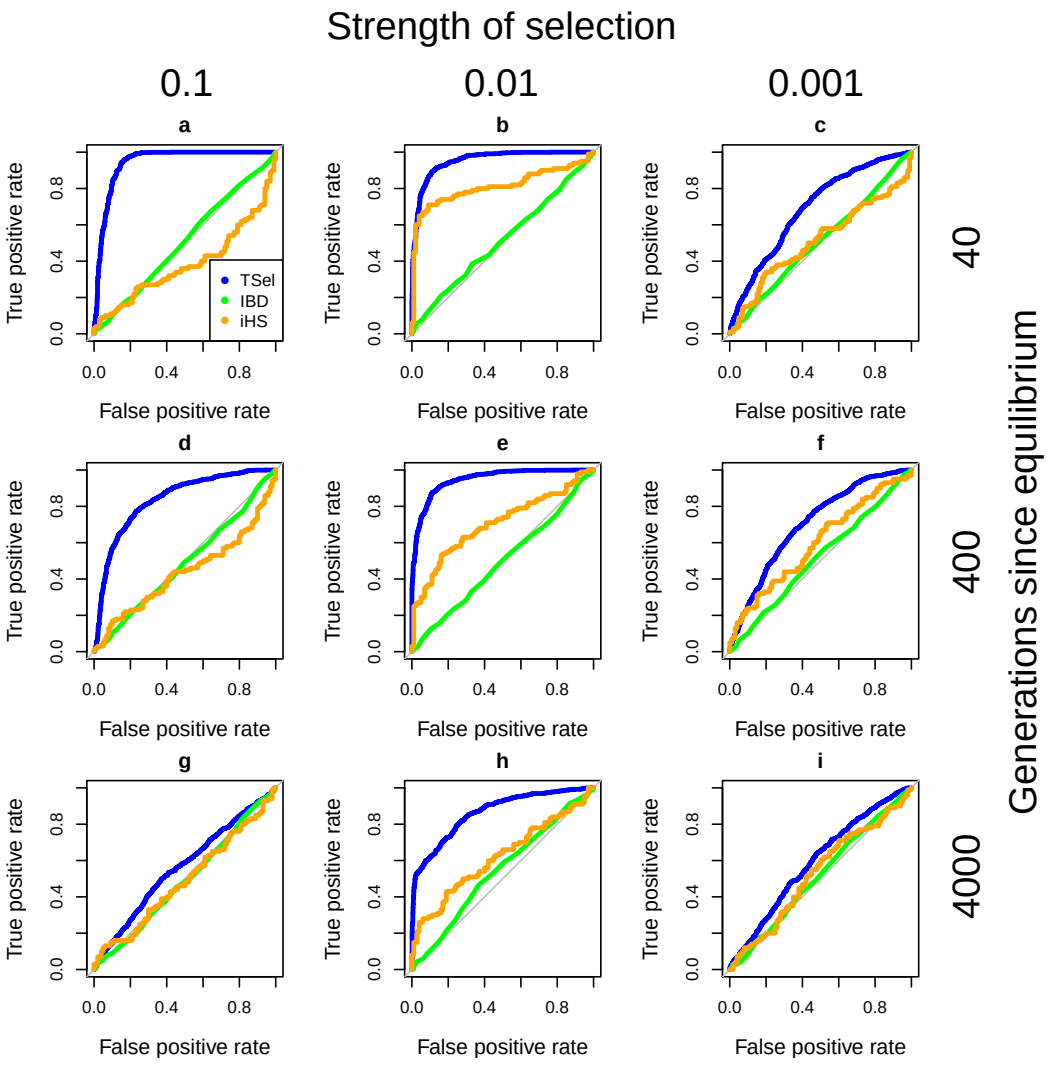


Figure 3.5: TSel performance on soft sweeps with an effective population size of 10,000

Performance is demonstrated via ROC curves. The initial frequency of the selected allele was set to 1%. The x-axis of the grid corresponds to the strength of selection and the y-axis corresponds to the time of sweep completion.

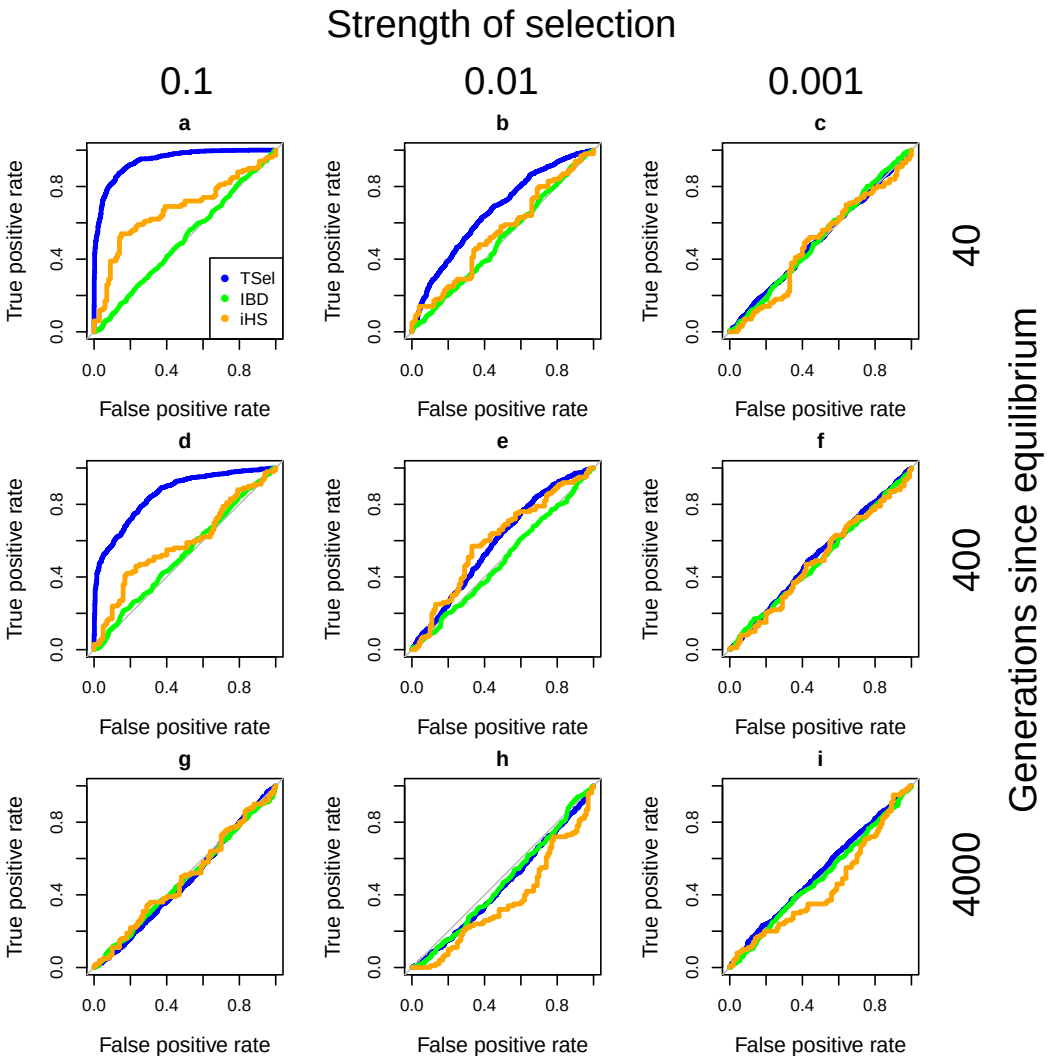


Figure 3.6: TSel performance on soft sweeps with an effective population size of 1,000

Performance is demonstrated via ROC curves. The initial frequency of the selected allele was set to 1%. The x-axis of the grid corresponds to the strength of selection and the y-axis corresponds to the time of sweep completion.

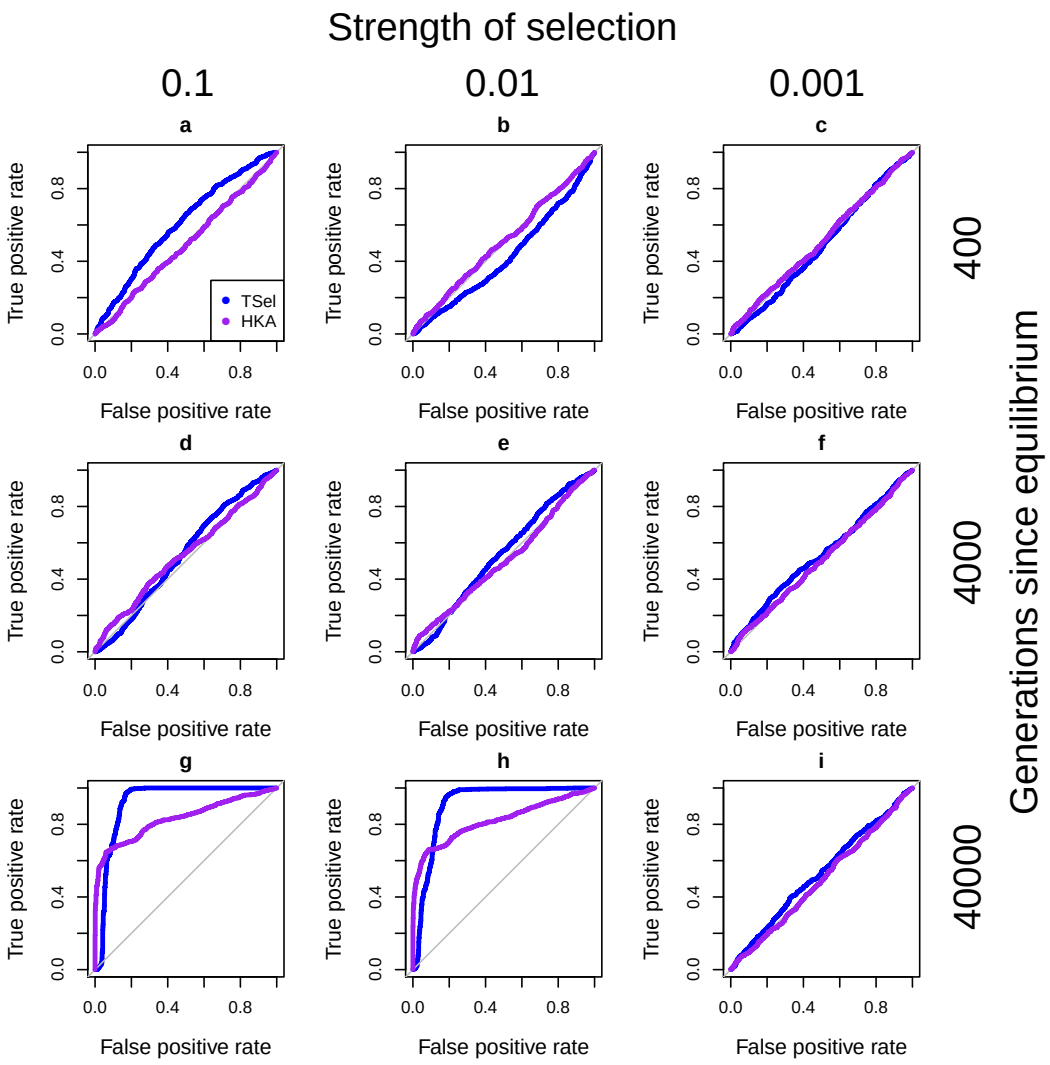


Figure 3.7: TSel performance on overdominance with an effective population size of 1,000

Performance is demonstrated via ROC curves. Selection began from one copy of the selected allele. The x-axis of the grid corresponds to the strength of selection and the y-axis corresponds to the time of the selected allele reached its equilibrium frequency of 0.5.

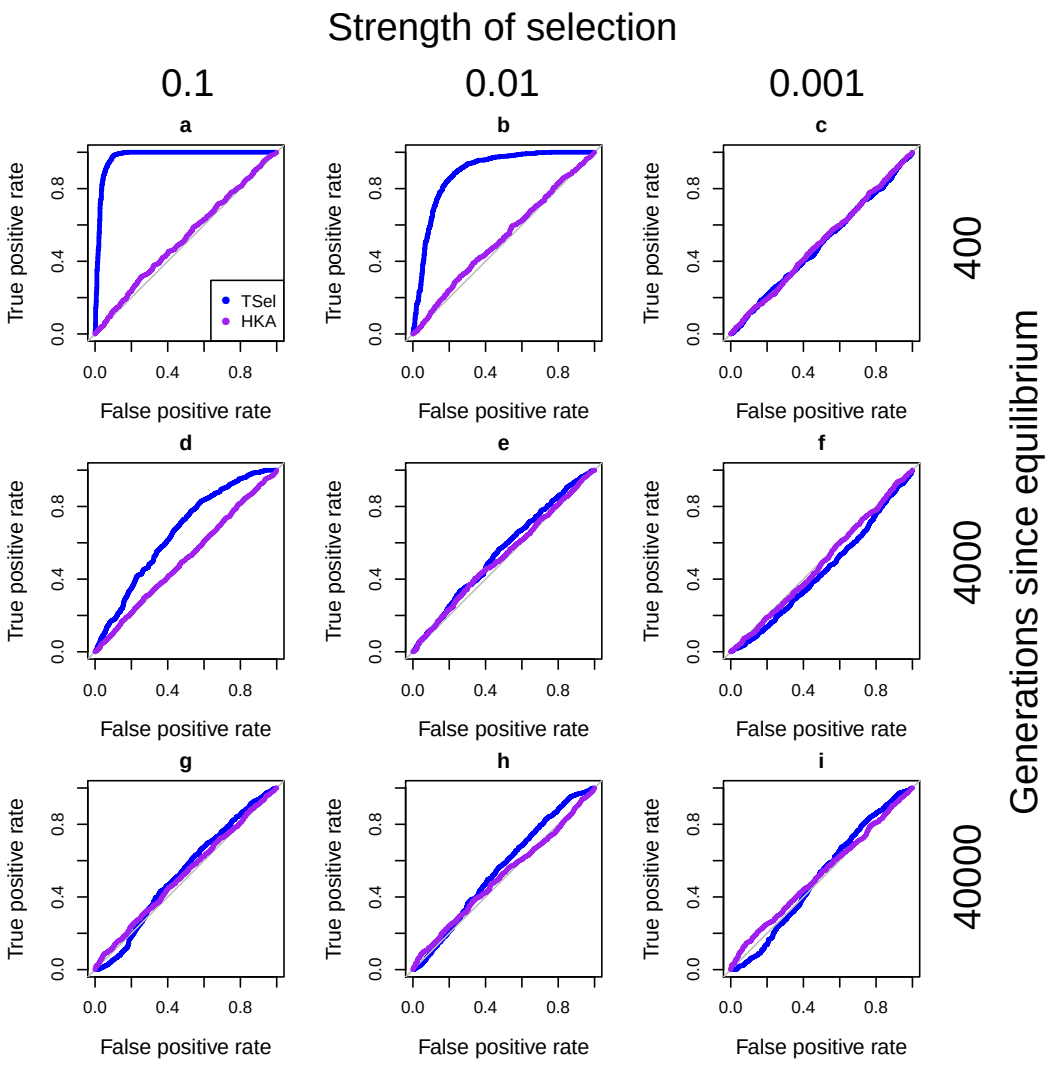


Figure 3.8: TSel performance on overdominance with an effective population size of 10,000

Performance is demonstrated via ROC curves. Selection began from one copy of the selected allele. The x-axis of the grid corresponds to the strength of selection and the y-axis corresponds to the time of the selected allele reached its equilibrium frequency.

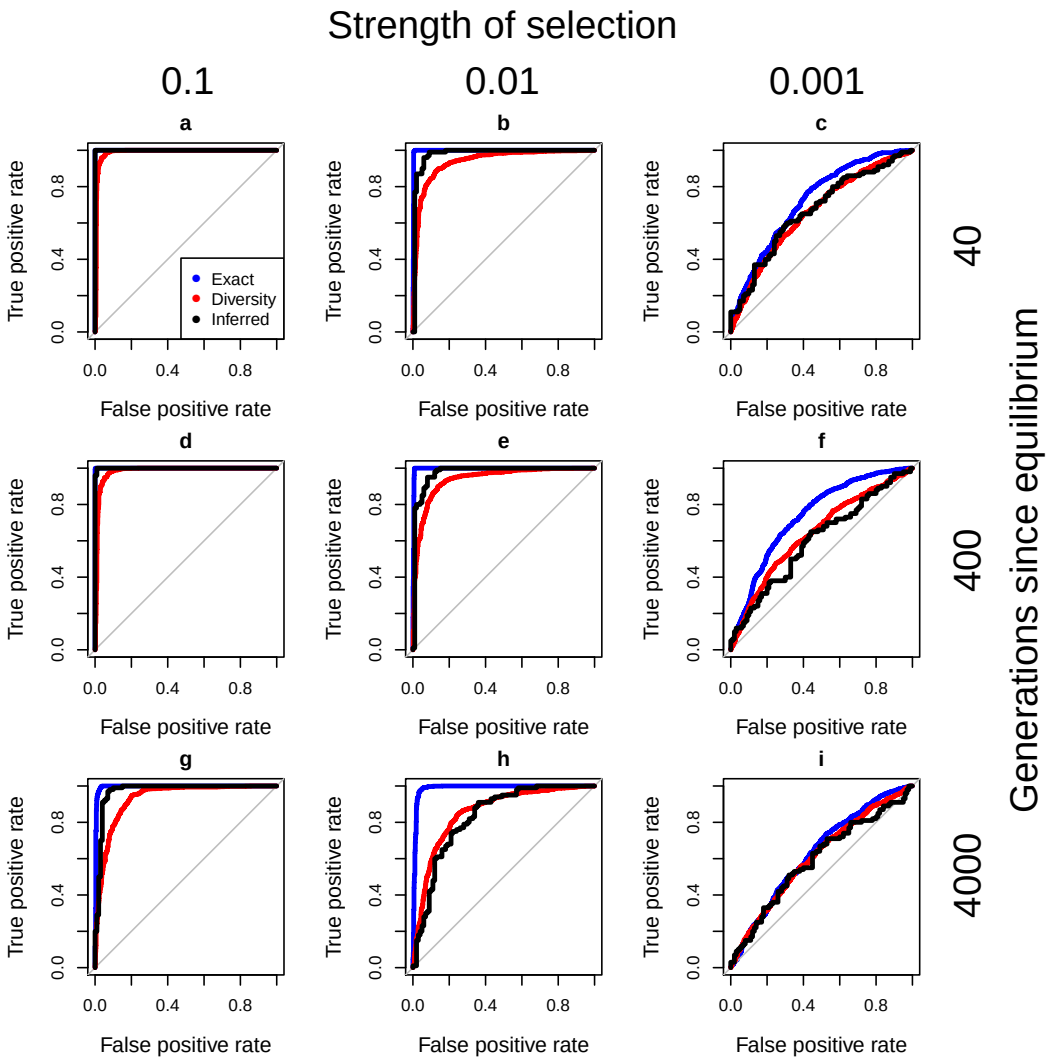


Figure 3.9: TSel performance using alternate feature sets

Performance is demonstrated via ROC curves on complete hard sweeps with an effective population size of 10,000. Performance is shown for TSel using features calculated from exact pairwise TMRCA distributions, PSMC-inferred pairwise TMRCA distributions, and genetic diversity. The x-axis of the grid corresponds to the strength of selection and the y-axis corresponds to the time of sweep completion.

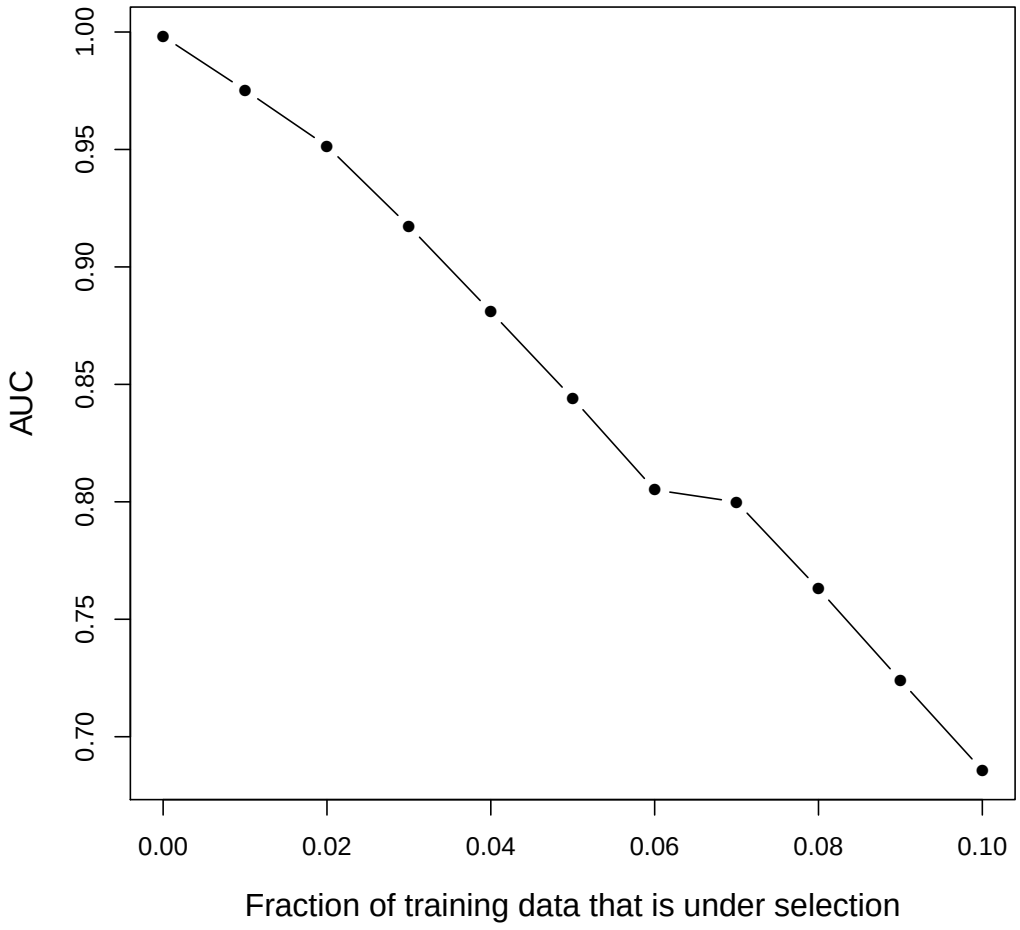


Figure 3.10: TSel performance when selected loci are not rare

Area under the curve (AUC) for ROC curves created for TSel performance when the data upon which the mean and covariance matrix for the Mahalanobis distance was calculated contains a certain fraction of selected data.

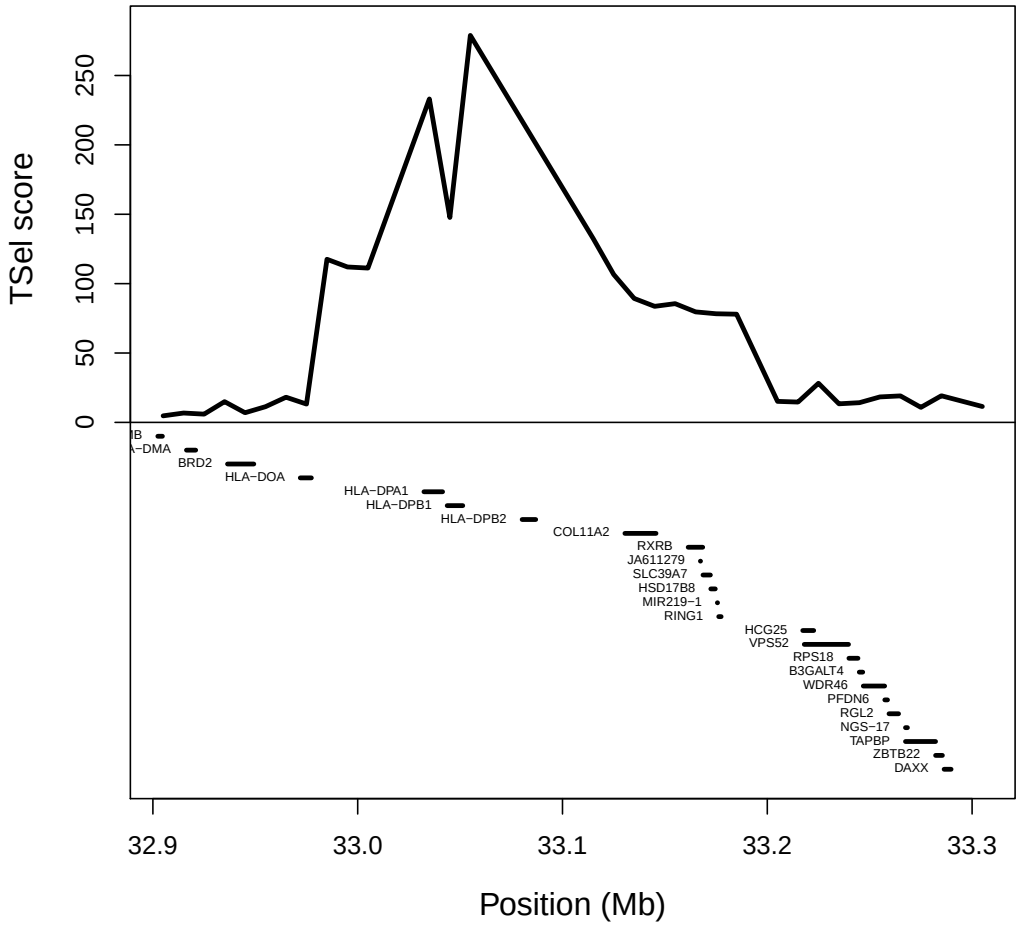


Figure 3.11: TSel replicates positive selection inference in real data

The figure displays part of the HLA region on chromosome 6. The top of the figure shows the median TSel score over each consecutive 10 kb window. The bottom of the figure shows genes that lie within the window. Grossman and colleagues used the method CMS to infer positive selection around the genes HLA-DPA1, HLA-DPB1, HLA-DPB2, a region that is also among the top 1% of TSel scoring regions.

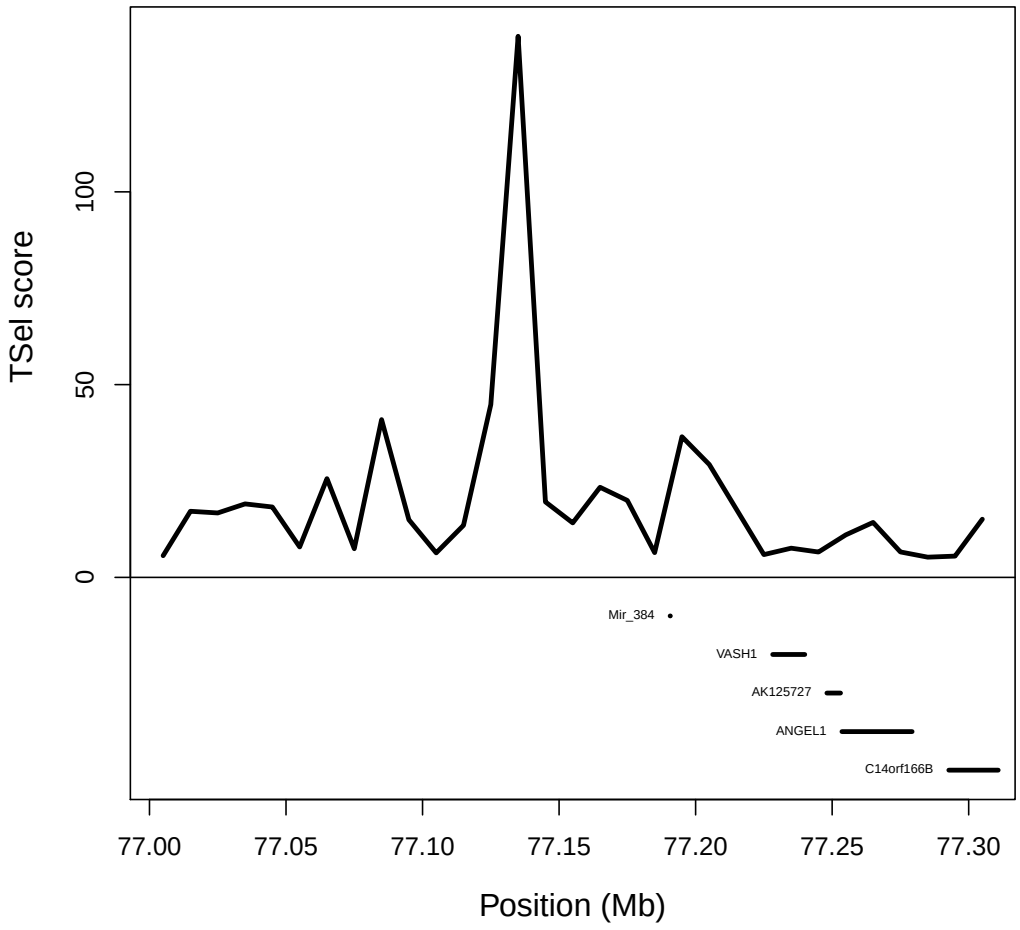


Figure 3.12: TSel replicates balancing selection inference in real data

The figure displays a region near the VASH1 gene on chromosome 14. The top of the figure shows the median TSel score over each consecutive 10 kb window. The bottom of the figure shows genes that lie within the window. Leffler and colleagues inferred ancient balancing selection near the VASH1 gene, a region that is also among the top 1% of TSel scoring regions.

CHAPTER 4

USING TIME TO MOST RECENT COMMON ANCESTOR IN ASSOCIATION MAPPING

Haley Hunter-Zinck and Andrew G. Clark

4.1 Abstract

Much of the genetic variance of complex traits has yet to explained. The missing heritability problem has many possible explanations, one of which is the lack of power of standard association methods to detect associations between a phenotype and rare variants. This limitation is particularly problematic in the genetic architectures involving allelic heterogeneity, and methods using relatedness metrics such as identity-by-descent (IBD) have been proposed as a possible solution. However, studies have shown that IBD mapping has limited applicability and using a more informative metric of genetic relatedness could help to improve performance of these methods. Here, we develop a new kernel for the sequence kernel association statistic (SKAT) test using time to most recent common ancestor (TMRCA). We then show that SKAT utilizing a kernel with pairwise TMRCA provides a more general framework for association mapping than using a kernel based on IBD, even if power is limited compared to other methods. We then go on to apply our new kernel on a real data set, looking for associations between X chromosome sequence and gene expression profiles in samples from the 1000 Genomes Project.

4.2 Introduction

Rare variant association studies have tried to address the missing heritability problem in discovering the genetic basis of disease as rare variants pose a challenge for traditional genome-wide scans, which lack the power to detect associations with variants of low frequency [77]. Previous methods and studies of coding regions have had notable but limited success in capturing the genetic components that contribute to phenotypic variation, and recent studies demonstrate that the vast majority of non-coding sequence is putatively functional [74]. Many rare variant methods are only applicable to a particular genetic phenotype model and have limited power to detect associations under a different framework [31]. Because of the wide range of genetic architectures that could underlie phenotypes, as well as the importance of non-coding genomic regions, there is a need to develop a unifying method applicable to whole genomes and to a large scope of genetic models.

Using statistics based on local relatedness between individuals has the potential to tackle this issue but has had limited progress so far. For example, identity-by-descent (IBD) mapping, using a signal of excess IBD, has been suggested as an effective metric of dealing with allelic heterogeneity at a single locus. However, simulation studies have revealed that IBD is only a useful metric for phenotype models that include an extreme amount of allelic heterogeneity [19]. Even though such a model may, in fact, be realistic under a demographic context with super-exponential growth, such as recently experienced in many human populations, using IBD does not provide an association test of wide applicability.

Although IBD does not provide a broad framework for association mapping, using another metric for local relatedness provides better results. Time to most recent common ancestor (TMRCA) is intimately related to IBD but a potentially more informative metric [16]. IBD inference methods often simplify relatedness to a binary call between a pair of individuals. Using pairwise TMRCA represents the relationship between two individuals as a set of continuous numbers rather than a single binary call. Combining the idea of using metrics of relatedness with a more informative metric of genetic relatedness in association studies has the potential to create a more unifying framework for association studies.

In order to incorporate the TMRCA metric, an association method that utilizes pairwise relatedness or similarity is necessary. The sequence kernel association statistic (SKAT) is one such method [78, 79]. SKAT has been shown to have more power than other rare variant methods for a combination of rare and common causal alleles using the linear weighted kernel based on the weighted product of genotypes. However, SKAT has limited power for other genetic phenotype models such as when causal variants are sampled with probability one over the minor allele frequency [31]. To try and increase the generality of SKAT, we can use SKAT with a kernel defined as a function of the local pairwise TMRCA between individuals, which could, at best, increase the scope of the association method or, at least, prove that using TMRCA is more informative than IBD in the association mapping context.

In the following study, we will develop and test a method using SKAT with a TMRCA kernel, which we will abbreviate TSKAT. We begin by testing TSKAT on a variety of simulated genetic models in a haploid context. We test TSKAT's performance over sample size, various demographic models, and genetic archi-

tectures. In order to exemplify the method in real data, we then run TSKAT on male X chromosome DNA samples with gene expression profiles from the 1000 Genomes Project [80, 81]. Overall, TSKAT shows limited performance on all models except associations with common variants in a population of constant size, but does outperform SKAT using an IBD kernel in all tested scenarios, providing evidence for the claim that TMRCA is a more informative metric of genetic relatedness than IBD. We then go on to discuss possible improvements to increase performance of TSKAT in a wider variety of demographic histories and genetic architectures.

4.3 Methods

4.3.1 TMRCA kernel

The SKAT method is a variance components test utilizing a kernel representing pairwise similarity between all samples, which can be represented by any positive semidefinite function [78, 79]. In order to incorporate metrics of genetic relatedness into the SKAT framework, we use a kernel derived from pairwise TMRCA values. Because pairwise TMRCA values represent difference rather than similarity, we used the negative log of the pairwise TMRCA value to represent the kernel entry between two individuals. The negative log satisfies the positive semidefinite function requirements because nearly identical individuals will have a small TMRCA which will result in a kernel value greater than zero; furthermore, all more distantly related individuals will have a TMRCA greater than that value and, therefore, a kernel value less than that value. We

then consolidate multiple non-recombining windows into a single 5000 bp window taking the median metric value over all non-recombining windows in the region.

4.3.2 Simulations

Using the simulator MSMS (version 3.2rc Build:147), we simulated data under a variety of scenarios to test the performance of the TMRCA kernel in an association framework [65]. We simulated a locus of 1 Mb in length and 100 replicates for each scenario. The recombination rate was set to 1×10^{-8} and the mutation rate was set to 1.1×10^{-8} [66]. To increase computational efficiency, recombination could only occur every 100 bp. In addition to simulated haplotypes, we also had the simulator output the coalescent trees, along with their respective scopes, for each non-recombining locus. MSMS provided a fast and flexible framework for simulating data to test TSKAT.

To test performance over different demographic models, we simulated two population histories. The first model represented a sample for a population of constant size at 10,000 individuals. The second model simulated a complex demographic history approximating a European human population [9]. We then tried two different genetic architectures to model the phenotype. In the first model, which we refer to as additive, only one high frequency casual locus affected the phenotype. The alternative model, which we refer to as heterogeneous, consisted of many different low frequency mutations, that all affected the phenotype. To select causal mutations, we first randomly chose a locus of 3 kb in length and selected 50% of the variants under 0.05 minor allele frequency

within that locus. Although the choice of locus size, fraction of variants, and minor allele frequency threshold are arbitrary, we later modified each of these parameters over an acceptable range to test the method's sensitivity to the choice of these parameters.

Following simulation procedures similar to those in the original SKAT paper, we simulated two covariates, one continuous variable sampled from a standard normal distribution and the second a binary variable being 0 or 1 with probability 0.5 [78]. The effect size of causal variants was calculated by taking the absolute value of the log of the minor allele frequency times a constant, where the constant was set to 0.4 to keep the values in a realistic range. We then defined each individual's phenotype using the genetic model, effect sizes, covariates, and standard normal error as follows:

$$y = \sum_{i=1}^{m} \beta_i x_i + 0.5c + 0.5b + \epsilon \tag{4.1}$$

where β_i is the effect size of causal variant i, x_i indicates whether the individual carries the causal variant i, c represents the continuous covariate value, b represents the binary covariate value, and ϵ represents the standard normal error. The resulting phenotypes were used in the association testing framework for each method tested.

4.3.3 TSKAT performance

We first analyzed TSKAT's performance for a sweep of sample sizes, analyzing the difference in performance for each sample size under the two genetic models

and the two demographic models. We chose sample sizes of 1000, 2000, 5000 haploid individuals and used the negative log of the pairwise TMRCA value to fill entries in SKAT's kernel. To assess performance over each sample size, we generated ROC curves for each sample size under each genetic model and demographic scenario [68].

We also assessed TSKAT's performance over different models of allelic heterogeneity when selecting causal alleles. Testing using the constant population size scenarios, we defined allelic heterogeneity over a broad sweep of parameters. First, we selected a frequency threshold of 0.01, 0.03, or 0.05 as a maximum allele frequency for causal SNPs. Next we chose the length of the causal locus as 1 kb, 3 kb, or 5 kb. And finally, we chose a fraction of variants under the frequency threshold and within the causal locus that would be considered causal. The fraction was set to 10%, 20%, or 50%. We then assessed the performance for each combination of parameters using the AUC.

4.3.4 Performance comparison

After assessing TSKAT's performance over sample size and heterogeneity models, we also compared performance to other association methods. Under both demographic and genetic models, we compared TSKAT's performance using the negative log TMRCA kernel with that of PLINK's Fisher's exact test, which we refer to as single marker analysis (SMA), SKAT with a weighted linear kernel, and SKAT with an IBD kernel [39]. Again, we assessed the relative performance of each method using ROC curves.

4.3.5 TSKAT application

In order to demonstrate the applicability of TSKAT in real data, we applied the method to the male X chromosome sequence data and X chromosome gene expression profiles collected on the 1000 Genomes Project dataset [80, 81]. Because of the hemizygous nature of the male X chromosome, this dataset provides a applicable sample for our method. In addition to being effectively haploid, the male X chromosome is also essentially phased, avoiding problems due to switch errors.

Before applying TSKAT, we first filtered the dataset extensively. We extracted male individuals for which we had both full genome sequence and gene expression values. For sites, we removed indels and two pseudoautosomal region at both ends of the X chromosome, and applied the 1000 Genomes Project strict masks to the sequence data. After pruning for linkage disequilibrium using the PLINK –indep-prune command and linkage disequilibrium correlation of 0.5, we ran principal components analysis on the X chromosome sequence data and used the first 10 principal components as covariates [39].

After preprocessing the data, we ran the pairwise sequentially Markovian coalescent (PSMC) method on all pairs of chromosomes [5]. Using TMRCA values output from PSMC, we constructed a TMRCA kernel using the negative log of the pairwise TMRCA value and took the median value for each pair to consolidate across multiple non-recombining windows. We used the resulting kernel for each 5000 bp window to generate a p-value via the SKAT method between each window and each X chromosome gene expression profile. After running TSKAT on the data, we then constructed a quantile-quantile plot to assess inflation and examined the significant hits after a Bonferroni correction

including all tests over all gene expression profiles and sequence windows.

4.4 Results

4.4.1 TSKAT performance

TSKAT performance by sample size is shown in Figure 4.1. Increases in sample size improve performance but the effects are relatively small. For example, increasing the sample size from 1000 haploid individuals to 5000 results in an increase of AUC of only 0.12 for a constant population size and the additive genetic model. The limited increase is consistent over both demographic scenarios and phenotypic models but even less for the heterogeneous phenotypic models. Increasing the sample size beyond 5000 will improve performance but at great computational cost because a linear increase in sample size means a quadratic increase in the number of pairs. Since performance does not dramatically improve between 1000 and 5000 individuals and computational time rises quadratically with the number of samples, we chose to use 1000 individuals for the remaining simulations.

However, the definition of the heterogeneous genetic model does greatly affect TSKAT's performance. Most notably, Figures 4.2, 4.3, and 4.4 demonstrate that TSKAT is sensitive to different combinations of causal locus length and fraction of causal variants given different thresholds of the minor allele frequency. For example, when the maximum minor allele frequency of causal variants is set to 3%, TSKAT obtains a maximum AUC of 0.6 for a causal locus length of 1 kb with 50% of the variants under that minor allele threshold selected. In

contrast, when the maximum minor allele frequency of causal variants is set to 5%, TSKAT performs best, with an AUC of 0.7, for a causal locus length of 5 kb with 50% of the variants under that minor allele threshold selected. Although, in general, the greater the number of causal variants, the greater TSKAT performance, these results demonstrate that patterns of performance rely on other factors as well, such as causal variant density.

4.4.2 Performance comparison

In order to compare TSKAT's performance to that of other association methods, we ran TSKAT along with several other methods on the same simulation scenarios and tested performance for each. Results are shown in Figure 4.5. The SKAT method with the IBD kernel performs poorly in all demographic scenarios and phenotypic models. SMA performs best for the additive phenotypic model but suffers for heterogeneous phenotypic model, where SKAT with the linear-weighted kernel does best. For the constant demographic history and the additive phenotypic model, TSKAT performance is below that of SMA but above that of SKAT with the linear-weighted kernel or IBD kernel. For the remaining scenarios, TSKAT is outperformed by both SMA and SKAT with the linear-weighted kernel. However, TSKAT outperforms SKAT with the IBD kernel in all tested demographic histories and phenotypic models, even with a complex demographic history and heterogeneous genetic model where TSKAT achieves an AUC of 0.60 and SKAT with the IBD kernel achieves only 0.54. Although SMA and SKAT with a linear-weighted kernel generally outperform TSKAT, TSKAT outperforms SKAT with an IBD kernel in all scenarios.

4.4.3 TSKAT application

We applied TSKAT to the 1000 genomes project male X chromosome sequence data and X chromosome gene expression profiles. After filtering for male individuals with both gene expression and whole-genome sequence data, we had a set of 121 individuals. Similarly for sites, after keeping only 121 individuals, applying the masks, removing indels, and removing the pseudoautosomal regions, we were left with 186,222 variant sites. Running TSKAT with the negative log TMRCA kernel with the first 10 principal components as covariates, the quantile-quantile plot, as shown in Figure 4.6, is below the expected distribution of p-values, indicating that we are over-correcting for population structure. Using a Bonferroni correction accounting for the number of tests over all gene expression values, we extract 270 significant associations over 10 unique gene expression profiles.

One of the significant associations is shown in Figure 4.7. This association occurs between variants located at approximately 136 Mb on chromosome X and the expression profile of the processed pseudogene KRT18P11 (ENSG00000215089). The associated region on chromosome X lies within 500 kb of genes ZIC3, a transcription factor, and ARHGEF6, involved in signaling pathways, indicating that variants in this region may have the ability to regulate the transcription of other genes, including KRT18P11. Although KRT18P11 is annotated as a pseudogene, trans eQTLs with pseudogene expression profiles have been discovered in previous studies [82].

4.5 Discussion

Using a TMRCA derived kernel provides a more general framework than IBD-mapping for associations in multiple demographic context and applicable to multiple phenotypic models. Although TSKAT does not out-perform SMA or SKAT with a linear-weighted kernel in all contexts, the performance improvements over that of IBD based kernel show promise. In addition, combining the results of SKAT methods with two different kernels, one derived from genotypes and another from TMRCA, could help generalize SKAT to both common variant and rare variant models of genetic architecture.

We could improve TSKAT's performance with respect to SMA or SKAT with a linear-weighted kernel in a variety of ways. Finding an optimal function for transforming the pairwise TMRCA values into the TMRCA-based kernel is crucial. The transformation may even vary depending on the underlying genetic architecture. Combining information with respect to singletons could also help improve TSKAT performance for rare variant associations. And finally, improving computational efficiency with respect to TMRCA kernel construction will allow users to increase sample size and, therefore, improve performance.

Although TMRCA inference is restricted at present to pairwise frameworks for all but very small sample sizes, the ultimate goal for future associations using TMRCA would be to use the full ancestral recombination graph. This step will necessitate a different methodology than that of SKAT, which is inherently pairwise, to associate complete trees with phenotypes of interest. Taking into account not only allelic similarity, but allelic similarity in the context of a detailed metric of common ancestry, will help to create more general methods of associ-

ation mapping. Creating more generic association methods will help shed light on the variety, type, and frequency of genetic architectures underlying complex disease and other phenotypes, a fundamental question of genetics.

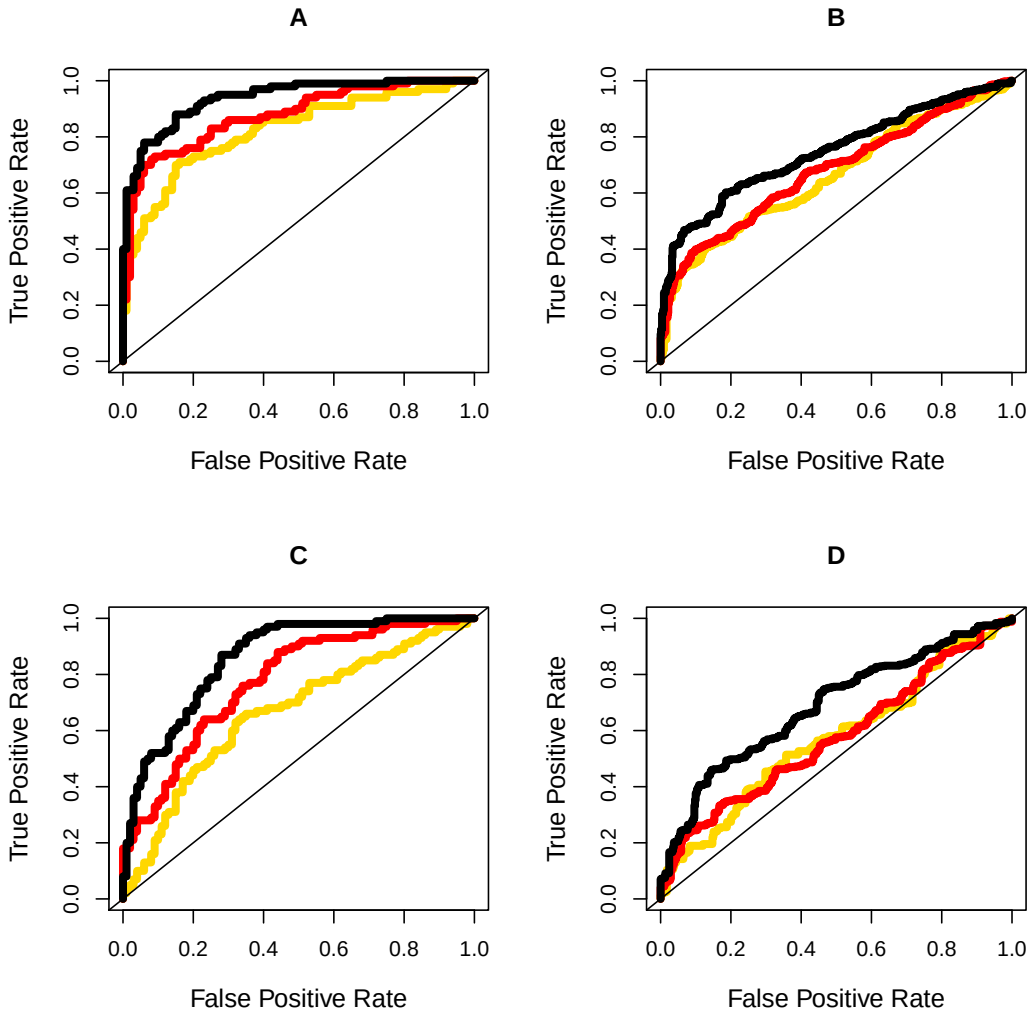


Figure 4.1: TSKAT performance by sample size

Rows represent the constant population size (A and B) and complex demography (C and D) and columns represent the additive (A and C) and heterogeneous (B and D) genetic models. Each subfigure contains three ROC curves representing a different number of sampled haploid individuals: 1000 (yellow), 2000 (red), and 5000 (black).

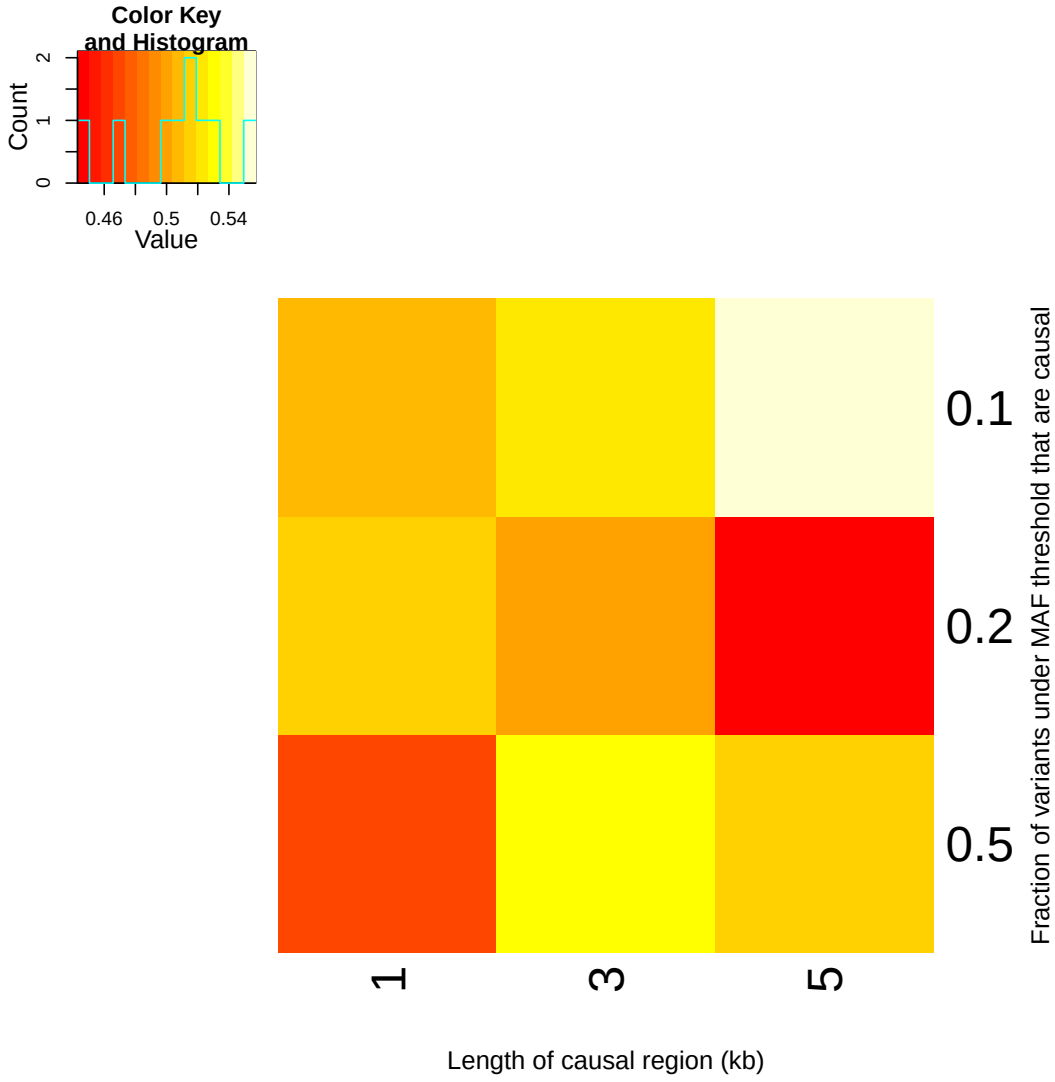


Figure 4.2: TSKAT performance for different definitions of allelic heterogeneity at 1% minor allele frequency.

Each cell of the heat map represents the AUC of TSKAT for a defined heterogeneity genetic model. All causal variants are below 1% minor allele frequency. The x-axis represents the length of causal region and the y-axis represents the fraction of variants in that region below 1% that are causal.

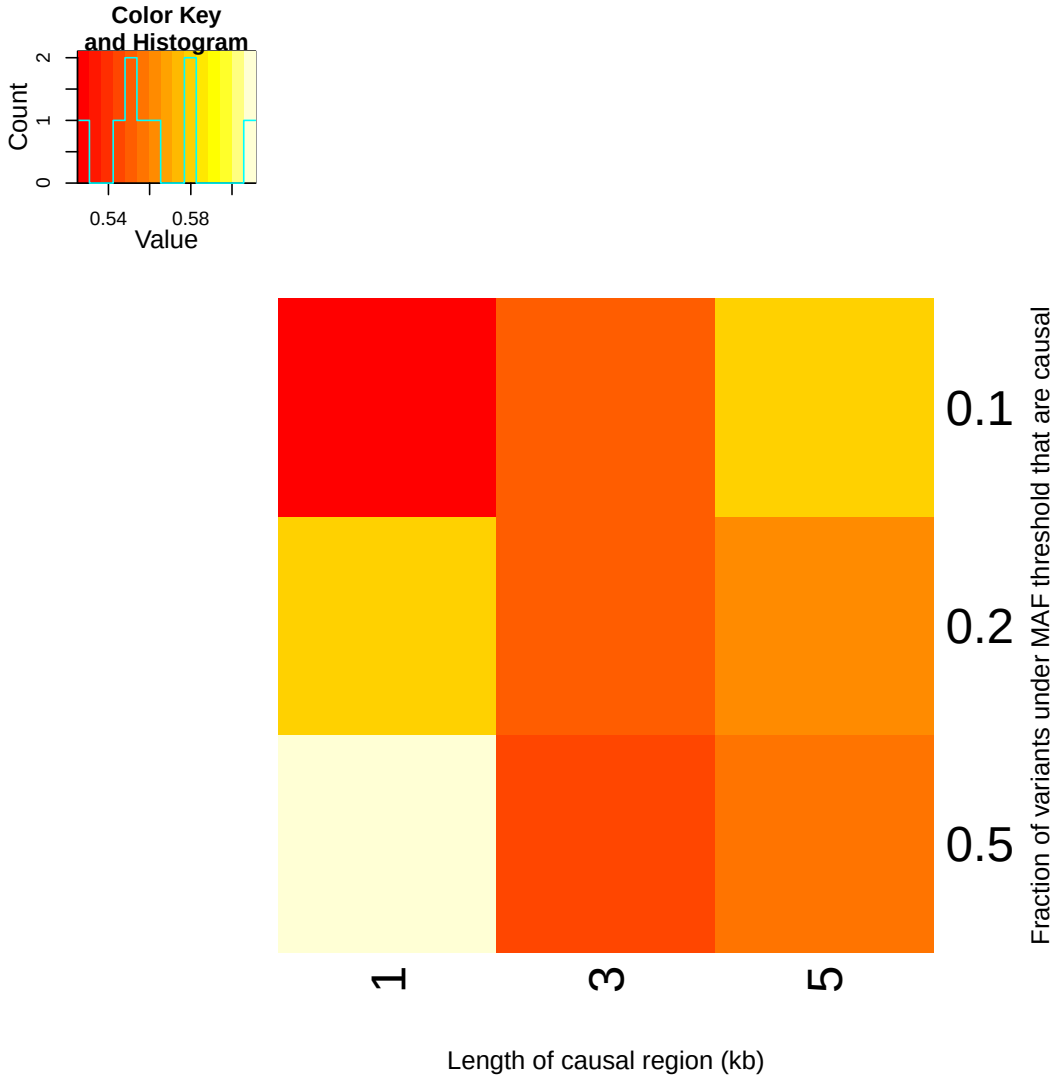


Figure 4.3: TSKAT performance for different definitions of allelic heterogeneity at 3% minor allele frequency.

Each cell of the heat map represents the AUC of TSKAT for a defined heterogeneity genetic model. All causal variants are below 3% minor allele frequency. The x-axis represents the length of causal region and the y-axis represents the fraction of variants in that region below 3% that are causal.

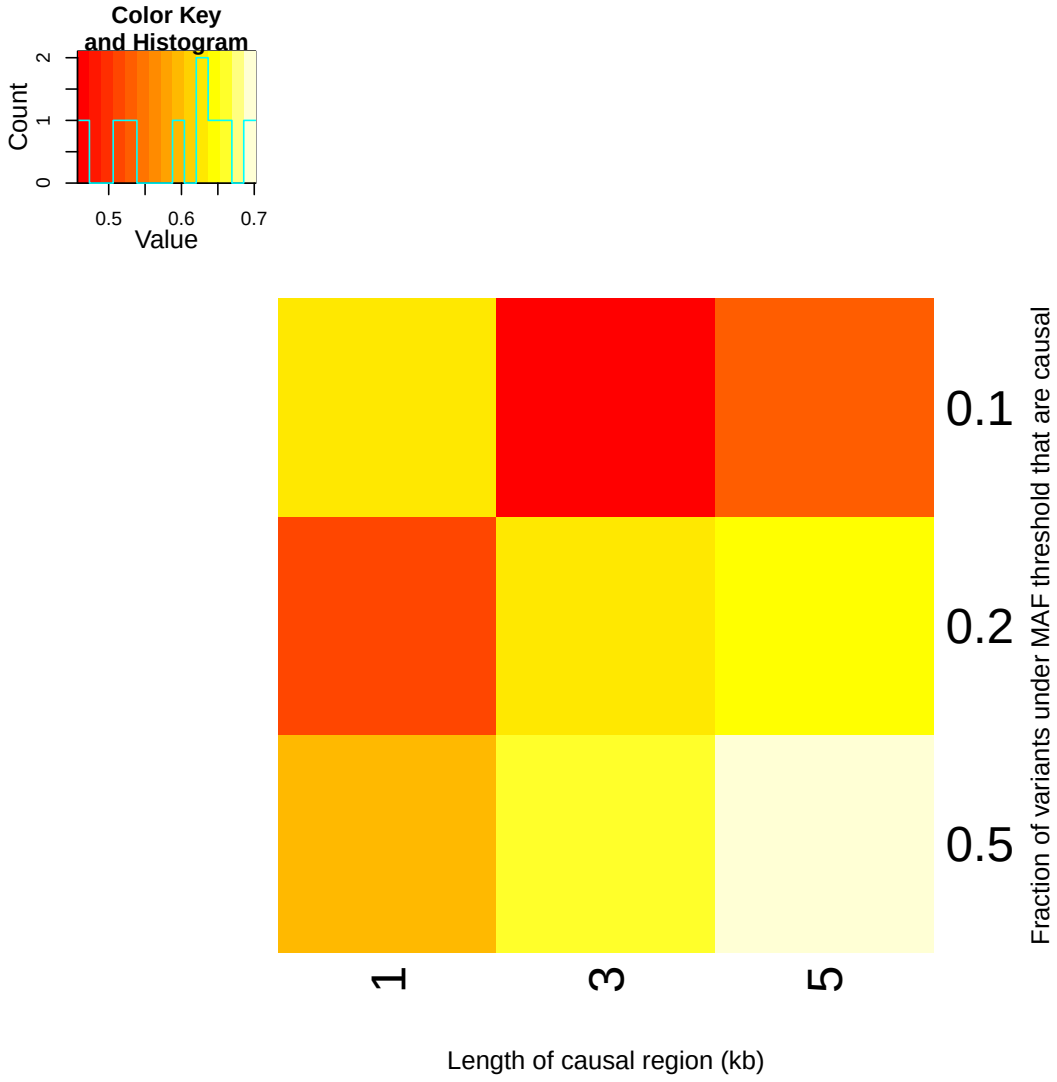


Figure 4.4: TSKAT performance for different definitions of allelic heterogeneity at 5% minor allele frequency.

Each cell of the heat map represents the AUC of TSKAT for a defined heterogeneity genetic model. All causal variants are below 5% minor allele frequency. The x-axis represents the length of causal region and the y-axis represents the fraction of variants in that region below 5% that are causal.

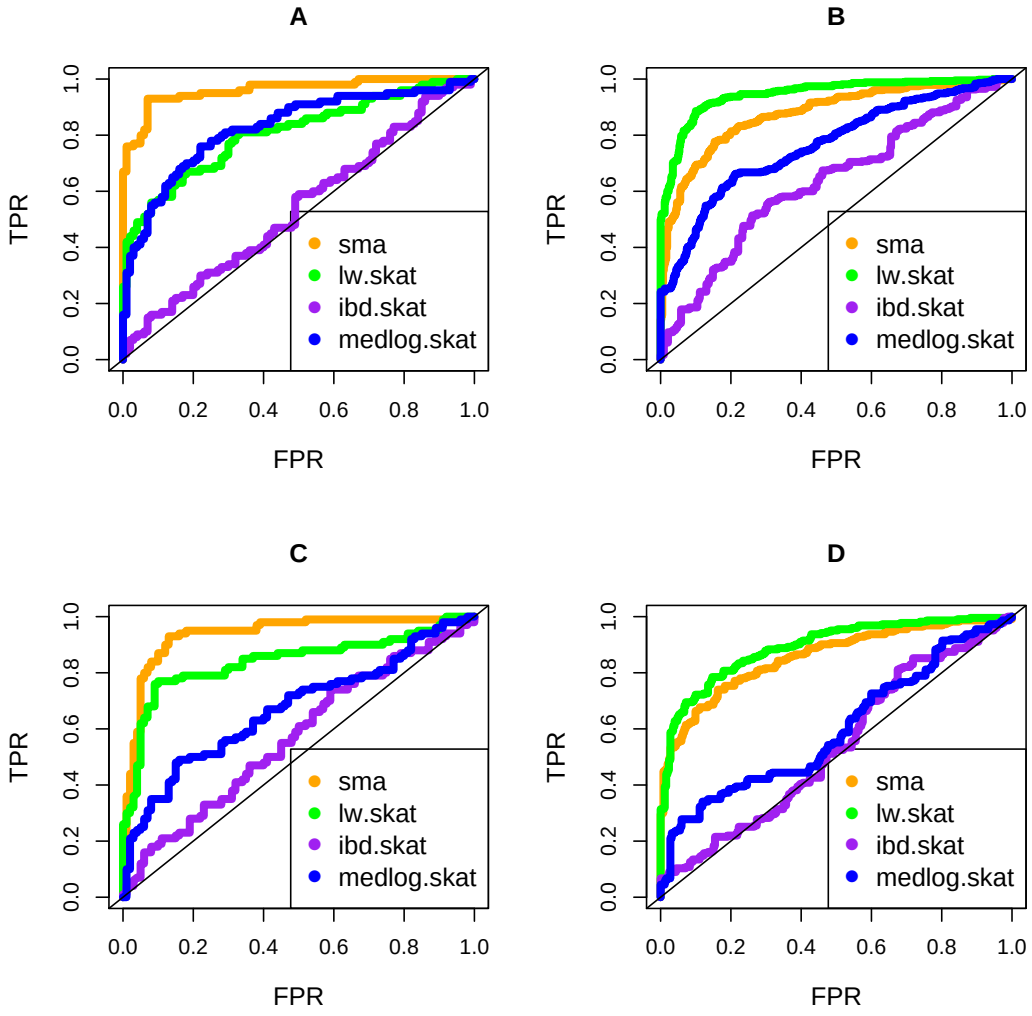


Figure 4.5: TSKAT performance as compared to other methods

TSKAT performance by sample sizeRows represent the constant population size (A and B) and complex demography (C and D) and columns represent the additive (A and C) and heterogeneous (B and D) genetic models. Each subfigure contains four ROC curves representing a different association method: single marker analysis (orange), SKAT with the linear-weighted kernel (green), SKAT with the IBD kernel (purple), and SKAT with the TMRCA kernel (blue).

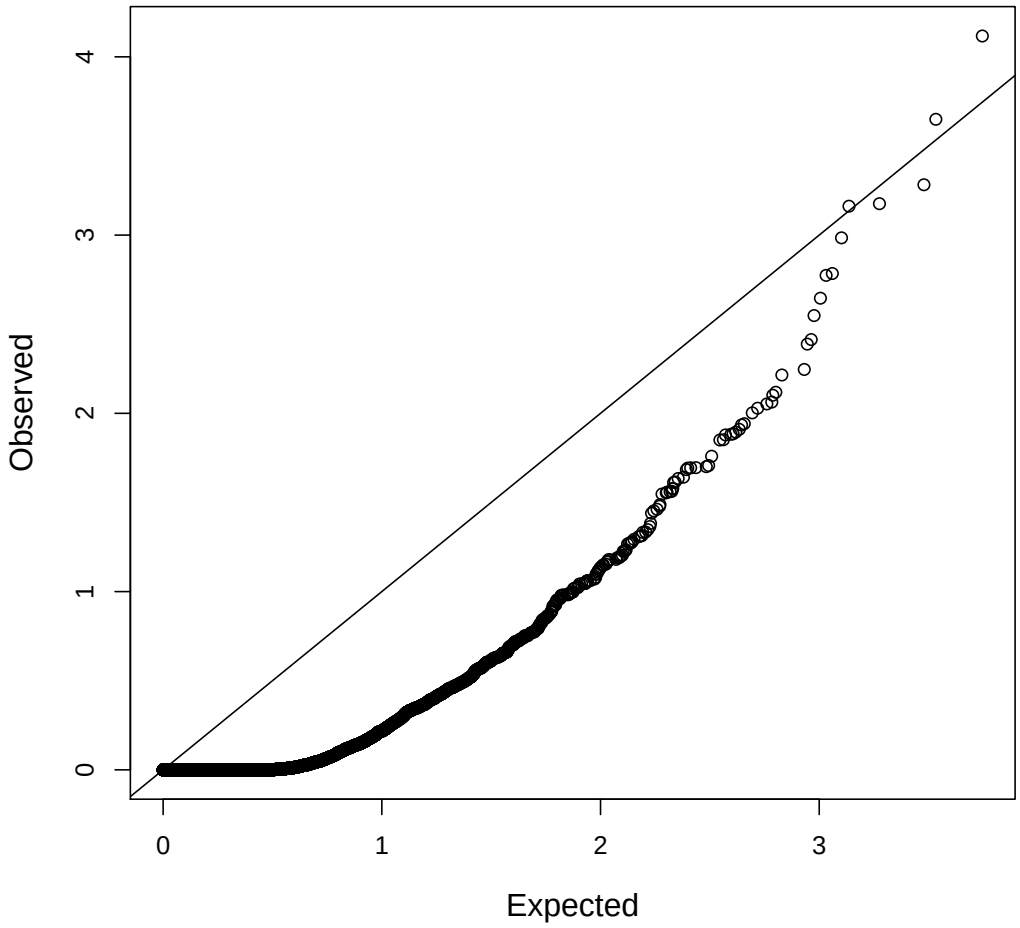


Figure 4.6: Quantile-quantile plot of significance values after applying TSKAT to associate X chromosome variants to X chromosome gene expression profiles.

The quantile-quantile plot compares the expected distribution of p-values, a uniform distribution, with the observed distribution of p-values for all gene expression profiles and sequence windows.

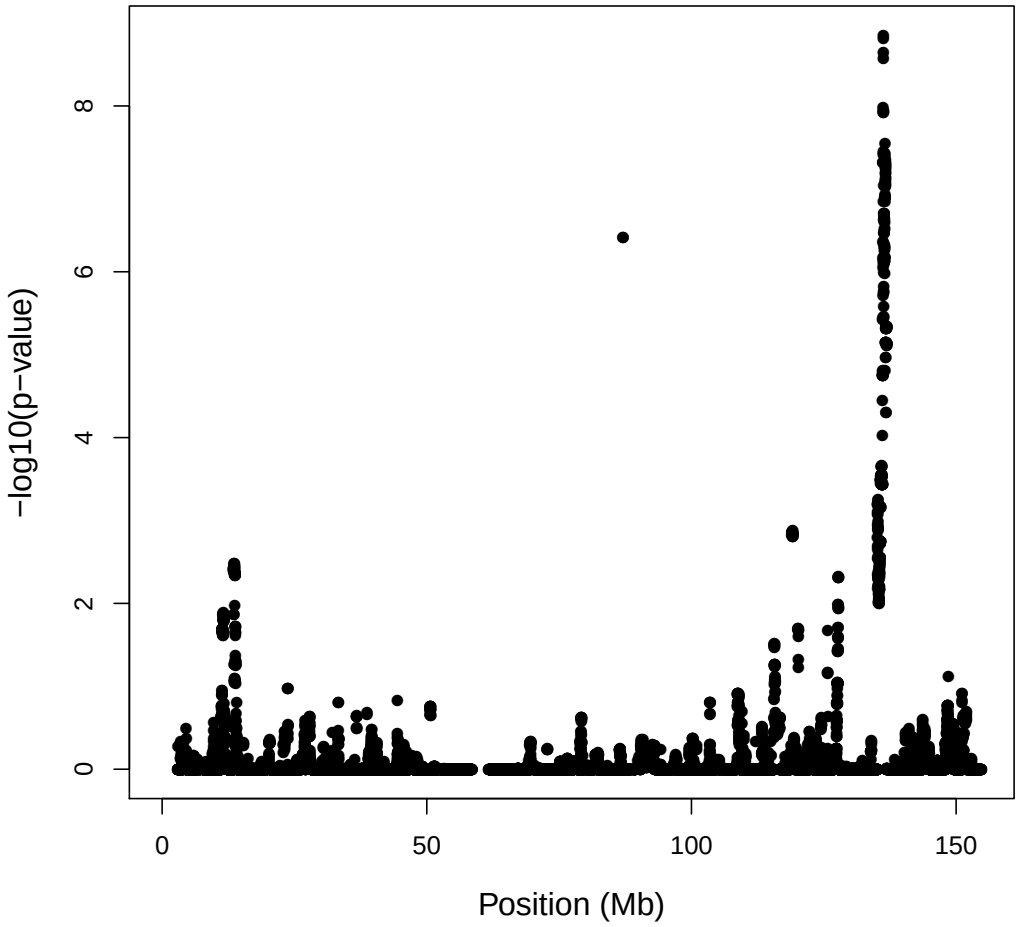


Figure 4.7: TSKAT association results of X chromosome variants with the gene expression profile of KRT18P11

The results of the TSKAT association of X chromosome sequence data with expression values for the processed pseudogene KRT18P11. The peak at approximately 136 Mb reaches significance after a Bonferroni correction for all tests over all gene expression profiles.

APPENDIX A

OVERSIZE CAPTIONS

A.1 Figure 2.3 caption

A. Principal components were calculated based on all HGDP populations and the Qatari data. Only Qatari data and HGDP Middle Eastern samples are graphed on this plot. Qatar1 clusters well with the other Middle Eastern populations, while Qatar2 creates a small cluster slightly removed from Qatar1 and the other Middle Eastern samples. Qatar3 does not form a definite cluster and is far removed from the main Middle Eastern cluster. **B**. Principal components were calculated only on Chinese and sub-Saharan African population samples. Qatari groups were then graphed on the plot using the principal components but were not used in the calculation of the principal components. The Qatar1 and Qatar2 groups cluster directly on top of the other Middle Eastern samples, which spread between the Asian and African groups. Qatar3 spreads between the Middle Eastern samples and the African samples. **C**. Principal components were calculated only on sub-Saharan African and Middle Eastern populations. Qatar groups then plotted onto these principal components. The Qatar3 group shows possible signs of admixture between the Middle Eastern cluster and the African population, while groups Qatar1 and Qatar2 cluster well with the other Middle Eastern populations. **D**. Principal components were calculated only on Chinese and Middle Eastern populations. Qatar groups were then plotted onto these principal components. The Qatar2 group shows a few individuals who demonstrate signs of admixture between the Middle Eastern samples and the Chinese samples but mostly cluster with the other Middle Easterners.

APPENDIX B

SUPPLEMENTARY MATERIALS: POPULATION GENETIC STRUCTURE OF THE PEOPLE OF QATAR

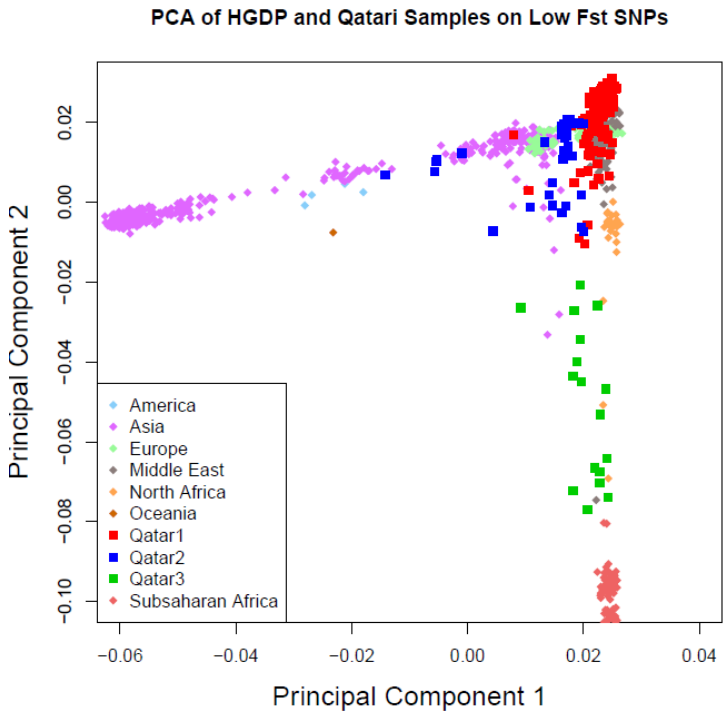


Figure B.1: Qatari and HGDP PCA analysis with low F_{ST} SNPs

We calculated F_{ST} for all SNPs between the Qatari samples and the HGDP European populations and selected a subset of SNPs that had FST below the mean. We repeated the PCA analysis on these SNPs and plotted the results. The resulting figure differs little from the original Figure 2.2 and, most importantly, the three clusters of the Qatari subgroups are still clearly visible as in the original figure. Therefore, although ascertainment bias may have more subtle effects on the analysis of the Qatari population sample, the discovery of the three Qatari subpopulations, as well as subsequent analysis on these three subgroups, seems robust to the choice of genotyped SNPs.

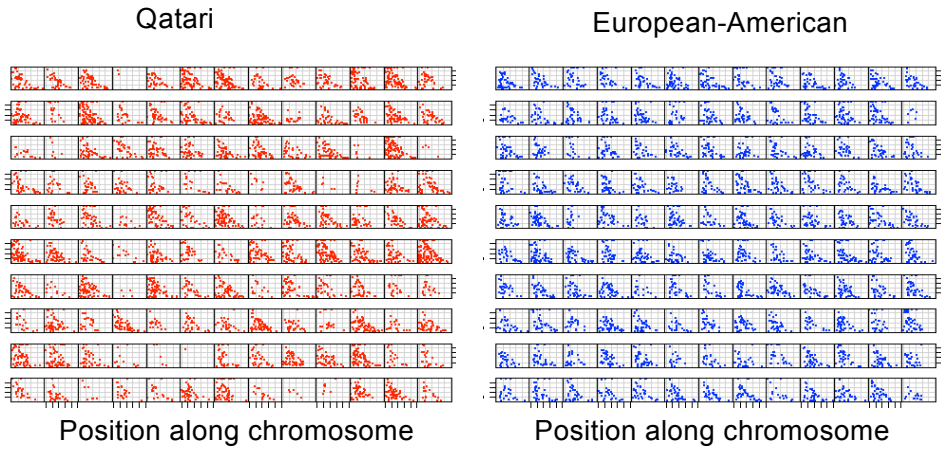


Figure B.2: Runs of homozygosity across Qatari and European-American samples.

On the left are plotted results from 130 individuals from the Qatari sample, and on the right are results from 130 individuals from the European-American population sample. Each cell of the figure is a different individual, and within each cell, the x axis is the position along a chromosome (0-250 Mbp) and the y axis is the chromosome number. Each colored segment represents a block of sequence in which SNP marker genotypes are homozygous. Note the great range in homozygosity among the Qatari compared to the European-Americans.

APPENDIX C

SUPPLEMENTARY MATERIALS: ABERRANT TIME TO MOST RECENT COMMON ANCESTOR AS A SIGNATURE OF NATURAL SELECTION

C.1 Details on MSMS commands

The MSMS commands used for all simulated experiments contained the following base command:

java -jar msms.jar *number_of_chromosomes number_of_replicates* -T -L -N N_e -t *theta* -r *rho* 10000 -oTPi 0.01 0.1 -oFP "#.#######"

For hard sweeps sweeps, we used the time independent-model of selection, adding the following switches to the base command:

-SAA *SAA* -SaA *SaA* -SF *time_sweep_completes* -Sp 0.5

with the appropriate selection coefficient and time. For all selection scenarios, the selected locus was placed in the middle of the simulated locus. Partial hard sweeps were simulated in a similar way but had an additional final frequency parameter:

-SAA *SAA* -SaA *SaA* -SF *time_sweep_completes final_frequency* -Sp 0.5

For soft sweeps, we used the time-dependent model of selection and conditioned on the survival of the allele throughout the forward simulations. We added the following switches to the base command:

-SAA *SAA* -SaA *SaA* -SI *time_selection_starts* 1 *initial_allele_frequency* -Sp 0.5 -SFC -oTrace

again, with the appropriate time, selection strength, and initial allele frequency for each selection scenario. Finally, for the overdominance scenarios, we used the time-dependent model of selection and also specified an outgroup sequence diverging from the main population at a time approximating chimp-human divergence. We added the following switches to the base command:

-Sc 0 1 0 *SaA* 0 -SI *time_selection_starts* 2 0.0001 0 -Sp 0.5 -I 2 100 1 -ej 8.125 1 2 -m 1 2 0 -SFC -oTrace

with each scenario's value for selection strength and start time.

C.2 Details on TSel R package

An implementation of TSel is available as the R package tsel and can be downloaded from the Clark lab website (`http://mbg.cornell.edu/research/clark-lab/`). The input file should have a header where the first two items are "chr position" and the remaining columns are labels for the features. Each row represents a locus where the first column specifies the locus' chromosome, the second column specifies the locus' position on that chromosome, and the remaining columns specify the respective feature values of that locus. This file should be input to the method getTselData(), which creates an object that is input to getTselScore(). The user may set both the quantile for inclusion of input features and the maximum threshold for correlation between features before removal. The method returns an object that contains the names of features included in the score calculation as well as the TSel score for each input locus.

The package also contains a summary and plotting method. The summary

prints out several summary statistics on the TSel results and the plotting method plots the TSel score by location in both a raw and smoothed format.

C.3 Filtering strategy for the Complete Genomics diversity panel

Sequencing and mapping errors result in spuriously deep estimates of TMRCA and potentially lead to false positive results for natural selection inference [5]. In order to avoid potential errors we masked genomic regions inferred to be problematic from analysis. To determine which regions could be problematic, we ran TSel on the unmasked data set and looked at the distribution of TSel scores across several region categories. From the UCSC genome browser, we downloaded tracks for segmental duplications, simple repeats, pseudogenes, the repeat masker, and the DAC blacklist. From the Complete Genomics diversity panel, we used files marking micro element insertions (MEIs), copy number variants (CNVs), structural variants, and unmapped regions. We also pulled down filters used in the INSIGHT method for CpG sites and recombination hotspots [83]. A plot of the median TSel score for each region category with standard error bars is shown in Supplementary Figure C.1. We ran a Mann-Whitney U test to determine if the distribution of scores sampled in each region category significantly differed from a random sample of TSel scores genome-wide. The filtered regions that significantly differ from the random sample are regions from the DAC blacklist, CNVs, CpG sites, segmental duplications, MEIs, pseudogenes, recombination hotspots, repeat masker, and unmapped regions. We included these regions in the final mask, which covers approximately 28% of

the genome in total and 34% of the transcribed regions genome-wide.

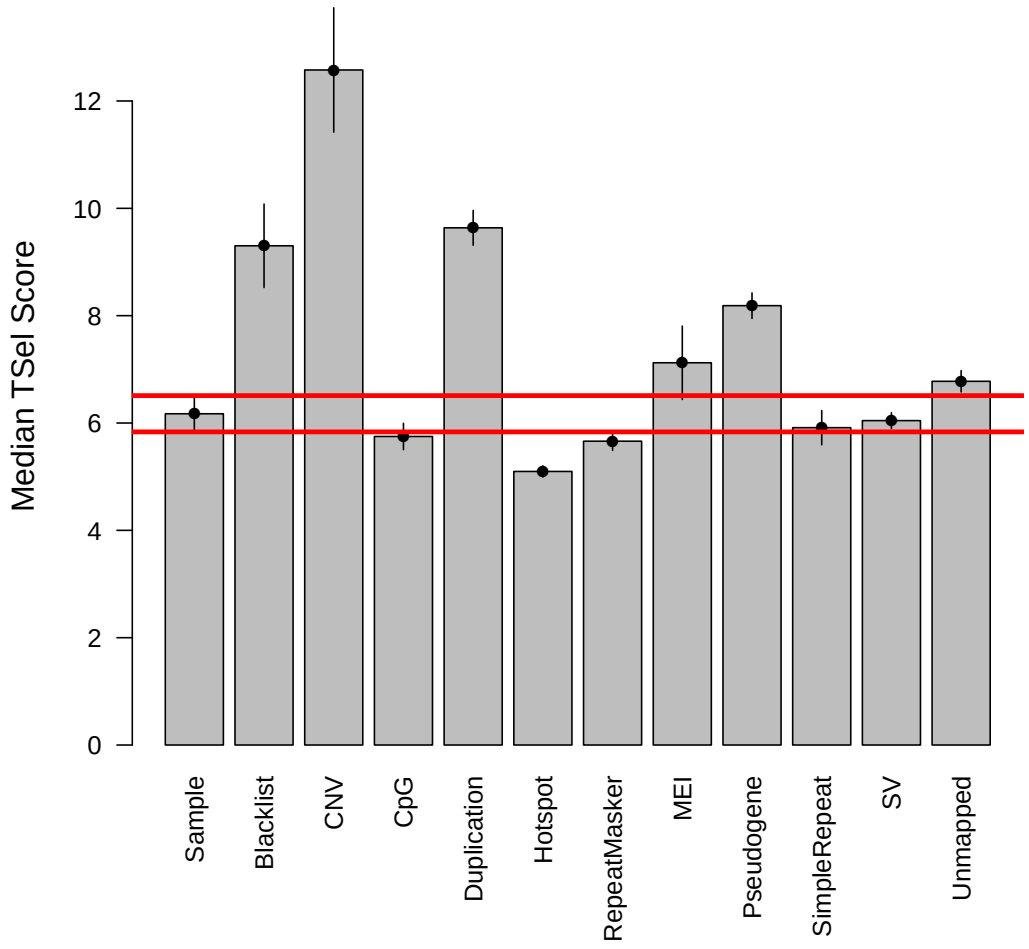
In addition to masking these regions, we also slightly modified the input procedure to PSMC by marking 100 bp windows as missing if they contained 20 or more missing base pairs. This procedure is more stringent than that employed by Li and Durbin in the original analysis, where they marked a window as missing only if 90 or more base pairs were missing from the 100 bp window. Although this approach works well for demographic applications, which summarizes data genome-wide, we wanted to employ a more stringent quality threshold for our analysis because the locus by locus values are relevant.

C.4 Assessing the effects of admixture

Although the TSel method assumes that detected deviations from the neutral genealogy are due to natural selection, forces other than selection, such as admixture, can also cause distortions of the local genealogies. However, we can distinguish between these two cases because admixture should be inconsistent with average neutral tree while selection could change the height and distances of the tree but not necessarily the consistency of the tree. Therefore, if admixture is a confounding factor, sites with high TSel scores should be more inconsistent with the neutral tree.

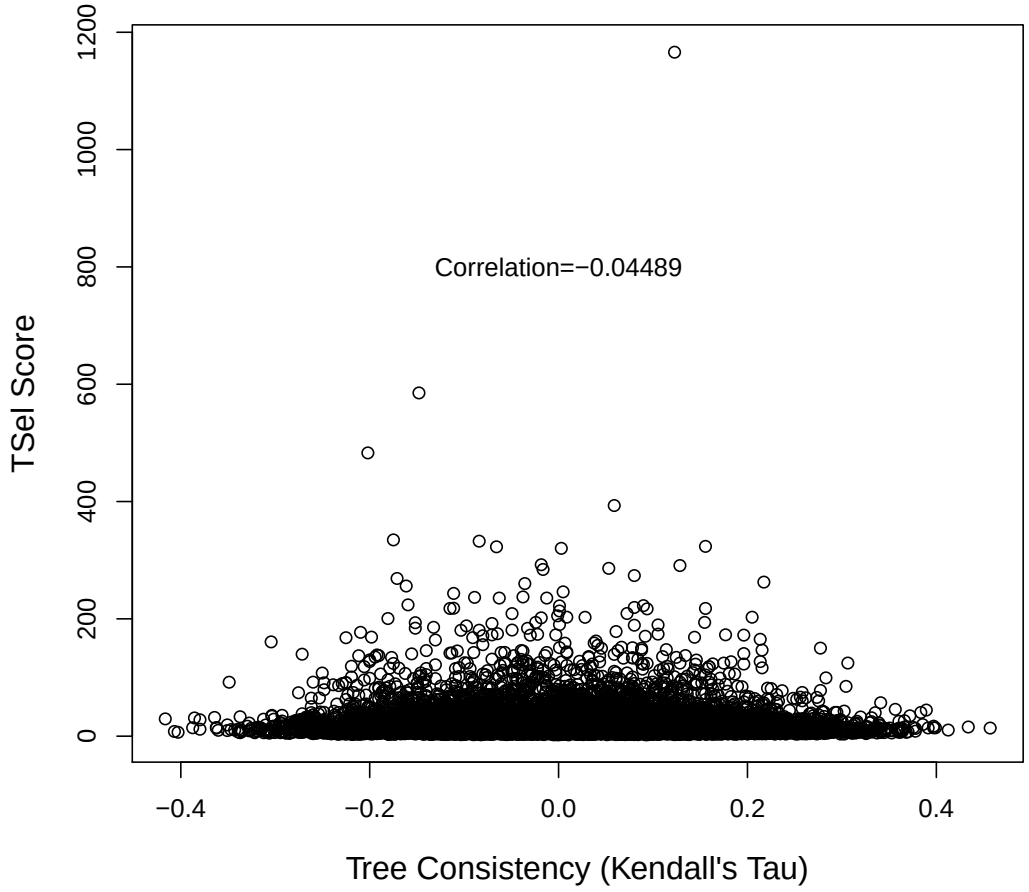
To assess whether high TSel scoring loci are more inconsistent with the average tree than expected, we looked at the ranking of pairwise TMRCA values between all samples and comparing each locus' ranking to the ranking of the average pairwise TMRCA values. We then calculated Kendall's τ rank correlation coefficient for each of these loci as a measure of tree consistency for that locus.

For a random sample of loci analyzed from the Complete Genomics diversity panel, we find little correlation between TSel score and tree inconsistency. Results are shown in Supplementary Figure C.2. Although this test does not completely rule out false positives due to admixture, TSel scores do not appear to be driven by tree inconsistencies.



Supplementary Figure C.1: TSel scores for potentially problematic region sets.

The sample column represents a random sample genome-wide and bars on each column indicate the median TSel score plus or minus the standard error. The red horizontal lines indicate plus or minus the standard error of the median of the random sample.



Supplementary Figure C.2: Assessing the effect of admixture on TSel scores.

The figure shows tree consistency, as measured by Kendall's τ, by TSel scores for a random sample of loci from the Complete Genomics diversity panel. The Pearson correlation coefficient is displayed in the center of the plot. The correlation between tree consistency and the TSel score is small, indicating that admixture is not driving high TSel Scores.

BIBLIOGRAPHY

[1] Wright S (1931) Evolution in Mendelian Populations. Genetics 16: 97–159.

[2] Novembre J, Johnson T, Bryc K, Kutalik Z (2008) Genes mirror geography within Europe. Nature 456: 98–101.

[3] McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genetics 5: e1000686.

[4] Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of Population Structure using Dense Haplotype Data. PLoS genetics 8: e1002453.

[5] Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493–6.

[6] Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740–3.

[7] Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337: 100–4.

[8] Palamara PF, Lencz T, Darvasi A, Pe'er I (2012) Length distributions of identity by descent reveal fine-scale demographic history. American journal of human genetics 91: 809–22.

[9] Tennessen Ja, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64–9.

[10] Albrechtsen A, Moltke I, Nielsen R (2010) Natural selection and the distribution of identity-by-descent in the human genome. Genetics 186: 295–308.

[11] Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, et al. (2013) Identifying recent adaptations in large-scale genomic data. Cell 152: 703–13.

[12] Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, et al. (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339: 1578–82.

[13] Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nature genetics 36: 512–7.

[14] Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nature reviews Genetics 8: 857–68.

[15] Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, et al. (2011) Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS genetics 7: e1001371.

[16] Thompson Ea (2013) Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. Genetics 194: 301–326.

[17] Han L, Abney M (2013) Using identity by descent estimation with dense genotype data to detect positive selection. European Journal of Human Genetics 21: 205–211.

[18] Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, et al. (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nature genetics 8: 380–386.

[19] Browning SR, Thompson Ea (2012) Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190: 1521–31.

[20] Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen F, et al. (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genetic epidemiology 33: 266–274.

[21] Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. Genome research 19: 318–26.

[22] Browning B, Browning S (2011) A Fast, Powerful Method for Detecting Identity by Descent. The American Journal of Human Genetics 88: 173–182.

[23] Su SY, Kasberger J, Baranzini S, Byerley W, Liao W, et al. (2012) Detection of identity by descent using next-generation whole genome sequencing data. BMC bioinformatics 13: 121.

[24] Browning BL, Browning SR (2013) Detecting identity by descent and estimating genotype error rates in sequence data. American journal of human genetics 93: 840–51.

[25] Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014) Genome-wide inference of ancestral recombination graphs. PLoS Genet 10: e1004342.

[26] Hunter-Zinck H, Musharoff S, Salit J, Al-Ali Ka, Chouchane L, et al. (2010) Population genetic structure of the people of Qatar. American journal of human genetics 87: 17–25.

[27] Sabeti P, Schaffner S, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614.

[28] Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome research 19: 711–22.

[29] Visscher PM, Brown Ma, McCarthy MI, Yang J (2012) Five years of GWAS discovery. American journal of human genetics 90: 7–24.

[30] Eyre-Walker A (2010) Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proceedings of the National Academy of Sciences 107: 1752–6.

[31] Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB (2012) The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS genetics 8: e1002496.

[32] Zeggini E, Scott L, Saxena R, Voight B (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nature Genetics 40: 638–645.

[33] Meyre D, Delplanque J, Chèvre J (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nature Genetics 41: 157–9.

[34] Larson M, Atwood L (2007) Framingham Heart Study 100K project:

genome-wide associations for cardiovascular disease outcomes. BMC Medical Genetics 8 Suppl 1: S5.

[35] Manolio T, Collins F, Cox N (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.

[36] Badii R, Bener A, Zirie M, Al-Rikabi A, Simsek M, et al. (2008) Lack of association between the Pro12Ala polymorphism of the PPAR-gamma 2 gene and type 2 diabetes mellitus in the Qatari consanguineous population. Acta Diabetologica 45: 15–21.

[37] Nagy S (2006) Making room for migrants, making sense of difference: Spatial and ideological expressions of social diversity in urban Qatar. Urban Studies 43: 119–137.

[38] Sandridge A (2010) Consanguinity in Qatar: knowledge, attitude and practice in a population born between 1946 and 1991. Journal of Biosocial Science 42: 59–82.

[39] Purcell S, Neale B, Todd-Brown K (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559–575.

[40] Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

[41] Li J, Absher D, Tang H, Southwick A (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–4.

[42] Hartl DL, Clark AG (2007) Principles of population genetics. Sinauer Associates, Inc.

[43] Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. PLoS Genetics 2: e190.

[44] Reich D, Price A, Patterson N (2008) Principal component analysis of genetic data. Nature Genetics 40: 491–492.

[45] Price A, Helgason A, Palsson S (2009) The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genetics 5: e1000505.

[46] Dray S, Dufour A (2007) The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software 22: 6.

[47] Gao H, Williamson S, Bustamante C (2007) A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176: 1635–1651.

[48] Rosenberg N, Pritchard J, Weber J (2002) Genetic structure of human populations. Science 2381: 2381–2385.

[49] Rosenberg N, Mahajan S, Ramachandran S (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genetics 1: e70.

[50] Tishkoff S, Reed F, Friedlaender F (2009) The genetic structure and history of Africans and African Americans. Science 324: 1035–1044.

[51] Reiner A, Ziv E, Lind D (2005) Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study. American Journal of Human Genetics 76: 463–77.

[52] Sheffield V, Stone E, Carmi R (1998) Use of isolated inbred human populations for identification of disease genes. Trends in Genetics 14: 391–6.

[53] Rice JA (1995) Mathematical Statistics and Data Analysis. Duxbury Press.

[54] Reich D, Cargill M, Bolk S, Ireland J (2001) Linkage disequilibrium in the human genome. Nature 411: 199–204.

[55] Abu-Amero K, Larruga J (2008) Mitochondrial DNA structure in the Arabian Peninsula. BMC Evolutionary Biology 8: 45.

[56] Abu-Amero K, Hellani A (2009) Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions. BMC Genetics 10: 59.

[57] Nelson M, Bryc K, King K, Indap A (2008) The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. American Journal of Human Genetics 83: 347–358.

[58] Auton A, Bryc K, Boyko A (2009) Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Research 19: 795–803.

[59] Reich D, Thangaraj K, Patterson N, Price A, Singh L (2009) Reconstructing Indian population history. Nature 461: 489–94.

[60] Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4: 99–111.

[61] Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327: 78–81.

[62] He X, Cai D, Niyogi P (2005) Laplacian score for feature selection. NIPS .

[63] Bishop CM (2006) Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc.

[64] Jr PJR, Diggle PJ (2001) geoR: a package for geostatistical analysis. R-NEWS 1: 14–18.

[65] Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26: 2064–5.

[66] Roach JC, Glusman G, Smit AFa, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636–9.

[67] Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43: 1031–4.

[68] Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940–1.

[69] Voight B, Kudaravalli S, Wen X, Pritchard J (2006) A map of recent positive selection in the human genome. PLoS Biol 4: 446–458.

[70] Grossman SR, Shlyakhter I, Shylakhter I, Karlsson EK, Byrne EH, et al. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327: 883–6.

[71] Gautier M, Vitalis R (2012) rehh: an R package to detect footprints of selec-

tion in genome-wide SNP data from haplotype structure. Bioinformatics 28: 1176–7.

[72] Hudson R, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.

[73] McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28: 495–501.

[74] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis Ca, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

[75] Sheehan S, Harris K, Song Y (2013) Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. Genetics 194: 647–662.

[76] Li H, Wiehe T (2013) Coalescent Tree Imbalance and a Simple Test for Selective Sweeps Based on Microsatellite Variation. PLoS Comput Biol 9: e1003060.

[77] Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, et al. (2014) Searching for missing heritability: Designing rare variant association studies. PNAS 111: E455–E464.

[78] Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. AJHG 89: 82–93.

[79] Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, et al. (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. AJHG 91: 1–14.

[80] 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

[81] Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506–511.

[82] Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, et al. (2012) Genetics of gene expression in primary immune cells identifies cell type specific master regulators and roles of HLA alleles. Nat Genet 44: 502–510.

[83] Gronau I, Arbiza L, Mohammed J, Siepel A (2013) Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Mol Biol Evol 30: 1159–71.